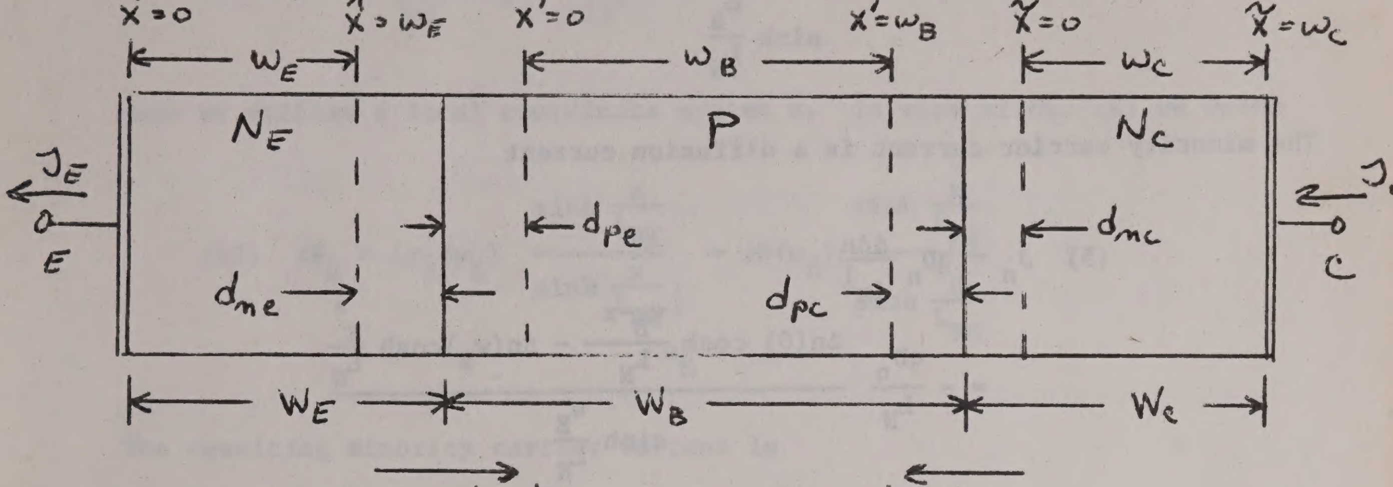
FORMAL DERIVATION OF JUNCTION TRANSISTOR EQUATIONS:

We will now consider the diffusion model of the junction transistor.



(a) Base

$\xrightarrow{\quad} V_{BE}{}^{+}$ $\qquad$ $+\xleftarrow{\quad} V_{BC}$

The minority carrier equation becomes

$$(1)\quad 0 = -\frac{\Delta n}{\tau_n} + D_n \frac{d^2 \Delta n}{dx'^2}$$

and

$$(2)\quad \Delta n = A_1 \varepsilon^{\frac{x'}{L_N}} + B_1 \varepsilon^{-\frac{x'}{L_N}}$$

we utilize a local coordinate system with its origin at the edge of the base-emitter depletion region. Then

$$(3)\quad \begin{bmatrix} \Delta n(0) \\ \Delta n(w_B) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \varepsilon^{\frac{w_B}{L_N}} & \varepsilon^{-\frac{w_B}{L_N}} \end{bmatrix} \begin{bmatrix} A_1 \\ B_1 \end{bmatrix}$$

or

$$\begin{bmatrix} A_1 \\ B_1 \end{bmatrix} = \frac{1}{2\sinh \frac{w_B}{L_N}} \begin{bmatrix} -\varepsilon^{-\frac{w_B}{L_N}} & 1 \\ \varepsilon^{\frac{w_B}{L_N}} & -1 \end{bmatrix} \begin{bmatrix} \Delta n(0) \\ \Delta n(w_B) \end{bmatrix}$$

Hence:

$$(4)\quad \Delta n = \frac{\Delta n(0)\ \sinh \frac{w_B - x'}{L_N} + \Delta n(w_B)\ \sinh \frac{x'}{L_N}}{\sinh \frac{w_B}{L_N}}$$

The minority carrier current is a diffusion current

$$(5)\quad J_n = qD_n \frac{d\Delta n}{dx^1}$$

$$= -\frac{qD_n}{L_N} \frac{\Delta n(0)\ \cosh \frac{w_B - x'}{L_N} - \Delta n(w_B) \cosh \frac{x'}{L_N}}{\sinh \frac{w_B}{L_N}}$$

the current at the two base edges becomes therefore:

$$(6)\quad J_n(x'=0) = -\frac{qD_n}{L_N} \Delta n(0)\ \coth \frac{w_B}{L_N} + \frac{qD_n}{L_N} \Delta n(w_B)\ \operatorname{csch} \frac{w_B}{L_N}$$

$$(7)\quad J_N(x'=w_B) = -\frac{qD_n}{L_N} \Delta n(0)\ \operatorname{csch} \frac{w_B}{L_N} + \frac{qD_n}{L_N} \Delta n(w_B)\ \coth \frac{w_B}{L_N}$$

We note that

$$(8)\quad \Delta n(0) = n_{p0} \left(\varepsilon^{-\frac{qv_{EB}}{kT}} - 1 \right) = n_{p0} \left(\varepsilon^{\frac{qv_{BE}}{kT}} - 1 \right)$$

$$(9)\quad \Delta n(w_B) = n_{p0} \left(\varepsilon^{-\frac{qv_{CB}}{kT}} - 1 \right) = n_{p0} \left(\varepsilon^{\frac{qv_{BC}}{kT}} - 1 \right)$$

and

$$(10)\quad w_B = W_B - d_{pe} - d_{pc}$$

(b) Emitter:

The emitter minority carrier equation is

$$(11)\quad 0 = -\frac{\Delta P_E}{\tau_{PE}} + D_P \frac{d^2 \Delta P_E}{d\hat{x}^2}$$

here we utilize a local coordinate system $\hat{x}$. In view of Eq. (4) we write

$$(12)\quad \Delta P_E = \Delta P_E(w_E) \frac{\sinh \frac{\hat{x}}{L_{PE}}}{\sinh \frac{\hat{x}}{L_{PE}}\Big|_{w_E}} = \Delta P(w_E) \frac{\sinh \frac{\hat{x}}{L_{PE}}}{\sinh \frac{w_E}{L_{PE}}}$$

The resulting minority carrier current is

$$(13)\quad J_{P.E} = -qD_{PE} \frac{d\Delta P_E}{d\hat{x}}$$

$$= -\frac{qD_{PE}}{L_{PE}} \Delta P_E(w_E) \frac{\cosh \frac{\hat{x}}{L_{P.E}}}{\sinh \frac{w_E}{L_{P.E}}} \quad .$$

The emitter minority carrier current at the edge of the depletion region becomes

$$(14)\quad J_{P E}(w_E) = -\frac{qD_{P,E}}{L_{P.E}} \Delta P_E(w_E) \coth \frac{w_E}{L_{P.E}}$$

where

$$(15)\quad P_E(w_E) \quad P_{E0}\left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right)$$

and

$$(16)\quad w_E = W_E - d_{nE} \quad .$$

(c) Emitter Terminal Current

If we neglect recombination in the emitter depletion region we note that

$$(17)\quad J_E = -J_{pE}(w_E) - J_N(x'=0) \quad .$$

The negative sign arises from the fact that the terminal current flows in the negative $\hat{x}$ and x' direction. By adding (6) and (14) we get

$$(18)\quad J_E = \left(\overbrace{\frac{qD_{PE}}{L_{PE}} P_{E0}\coth\frac{w_E}{L_{PE}}}^{\text{diff in emitter}} + \overbrace{\frac{qD_n}{L_n} n_{p0}\coth\frac{w_B}{L_N}}^{\text{diff in base}}\right)\underbrace{\left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right)}_{\text{due to BE boundary}}$$

$$- \underbrace{\frac{qD_n}{L_N} n_{P0}\operatorname{csch}\frac{w_B}{L_N}}_{\text{diff in base}}\underbrace{\left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right)}_{\text{due to BC boundary}} .$$

(d) Collector

The collector minority carrier equation is

$$(19)\quad 0 = -\frac{\Delta P_C}{\tau_{PC}} + D_{PC}\frac{d^2\Delta P_C}{d\tilde{x}^2} .$$

Here again we are using a shifted coordinate system. In view of Eq. (4) we write

$$(20)\quad \Delta P_C = \Delta P_C(0)\frac{\sinh\frac{w_C-\tilde{x}}{L_{PC}}}{\sinh\frac{w_C}{L_{PC}}} .$$

The resulting minority carrier current becomes

$$J_{PC} = -qD_{PC}\frac{d\Delta P_C}{d\tilde{x}}$$

$$(21)\quad J_{pC} = \frac{qD_{PC}}{L_{PC}}\frac{\cosh\frac{w_C-\tilde{x}}{L_{PC}}}{\sinh\frac{w_C}{L_{PC}}}\Delta P_C(0) .$$

Here

$$(22)\quad \Delta P_C(0) = P_{C0}\left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right)$$

$$(23)\quad w_C = W_C - d_{nC} .$$

The current at the boundary becomes

$$(24)\quad J_{PC}(\tilde{x}=0) = \frac{qD_{PC}}{L_{PC}}\coth\frac{w_C}{L_{PC}}\,\Delta P_C(0) .$$

(e) Collector Terminal Current

We again neglect recombination in the space-charge region. Then

$$(25)\quad -J_C = J_{PC}(\tilde{x}=0) + J_N(x'=w_B)\ .$$

Combining (24) and (7) yields

$$(26)\quad -J_C = \left(\overbrace{\frac{qD_{PC}}{L_{PC}}\,P_{C0}\,\coth\frac{w_C}{L_{PC}}}^{\text{diff in collector}} + \overbrace{\frac{qD_n}{L_N}\,n_{P0}\,\coth\frac{w_B}{L_N}}^{\text{diff in base}}\right)\underset{\uparrow\ \text{base collector boundary}}{\left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right)}$$

$$- \underbrace{\frac{qD_n}{L_N}\,n_{P0}\,\operatorname{csch}}_{\text{diff in base}}\frac{w_B}{L_N}\underset{\uparrow\ \text{emitter base boundary}}{\left(\varepsilon^{\frac{qV_{BC}\ \text{— VBE}}{kT}} - 1\right)}.$$

(f) Diode Model

We combine equations (18) and (26) to obtain:

$$(27)\quad \begin{bmatrix} J_E \\ \\ -J_C \end{bmatrix} = \begin{bmatrix} \frac{qD_{PE}}{L_{PE}}\,P_{E0}\,\coth\left(\frac{W_E-d_{nE}}{L_{PE}}\right) + \frac{qD_n}{L_N}\,n_{P0}\,\coth\left(\frac{W_B-d_{PE}-d_{PC}}{L_N}\right) \\ \\ -\frac{qD_n}{L_N}\,n_{P0}\,\operatorname{csch}\left(\frac{W_B-d_{PE}-d_{PC}}{L_N}\right)\quad \frac{qD_{PC}}{L_{PC}}\,P_{C0}\,\coth\frac{W_C-d_{nC}}{L_{PC}} \end{bmatrix}$$

$$\left.\begin{matrix} -\frac{qD_n}{L_N}\,n_{P0}\,\operatorname{csch}\frac{W_B-d_{PE}-d_{PC}}{L_N} \\ \\ +\frac{qD_n}{L_N}\,n_{P0}\,\coth\frac{W_B-d_{PE}-d_{PC}}{L_N} \end{matrix}\right]\ \left.\begin{matrix} \left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right)\ \text{— VBE} \\ \\ \left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right) \end{matrix}\right]$$

This equation is the diffusion model description of the bi-polar transistor. Examination of Eq. (27) yields the conclusion that the element m_{11} is the base-emitter diode saturation current. The element m_{22} is the base-collector saturation current density. We write

$$(28)\quad \begin{bmatrix} J_E \\ J_C \end{bmatrix} = \begin{bmatrix} J_{SE} & -a_{12} \\ -a_{21} & J_{SC} \end{bmatrix}\begin{bmatrix} \left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right) \\ \left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right) \end{bmatrix}$$

Here

$$a_{12} = a_{21} = \frac{qD_n}{L_n} n_{P0} \operatorname{csch} \frac{W_B - d_{PE} - d_{PC}}{L_N}$$

The coupling terms a_{12} and a_{21} approach zero for large values of W_B. The system would then reduce to two simple diodes. Thus, in a sense a transistor is a doulbe diode with one common region which is short. Suppose we measure the effect on J_E of changing V_{BE}:

$$\left.\frac{\partial J_E}{\partial V_{BE}}\right|_{\Delta V_{BC}=0} = \frac{q}{kT} \varepsilon^{\frac{qV_{BE}}{kT}} J_{SE}$$

and the effect on the collector current becomes

$$\left.\frac{\partial J_C}{\partial V_{BE}}\right|_{\Delta V_{BC}=0} = + \frac{q}{kT} \varepsilon^{\frac{qV_{BE}}{kT}} a_{21}$$

If we now define the current amplification factor α_F by

$$(29) \quad \alpha_F \equiv \frac{\dfrac{\partial J_C}{\partial V_{BE}}}{\dfrac{\partial J_E}{\partial V_{BE}}} = + \frac{a_{21}}{J_{SE}}$$

then $$+a_{21} = \alpha_F J_{SE} .$$

A similar argument yields

$$(30) \quad \alpha_R \equiv \frac{\dfrac{\partial J_E}{\partial V_{BC}}}{\dfrac{\partial J_C}{\partial V_{BC}}} = + \frac{a_{12}}{J_{SC}}$$

$$+a_{12} = \alpha_R J_{SC}$$

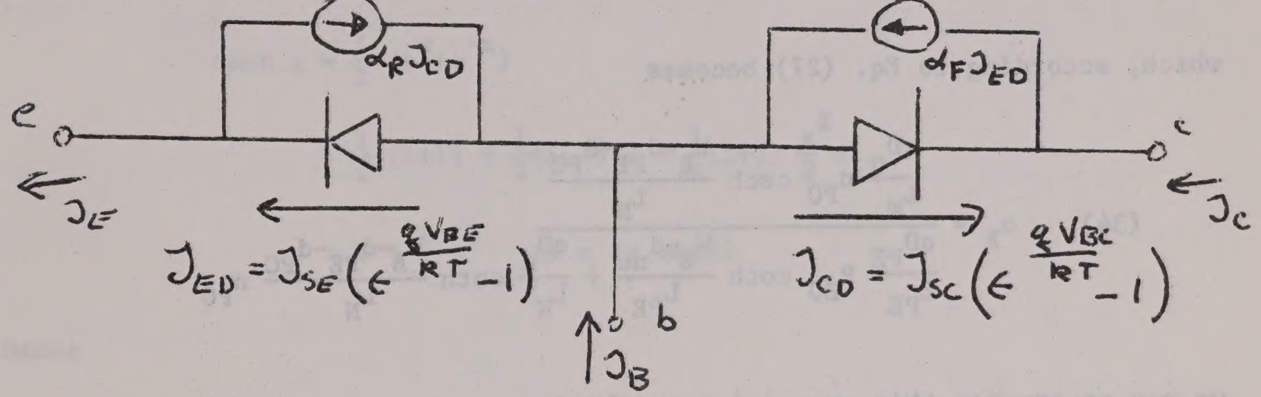
and, since $a_{12} = a_{21}$

$$(31) \quad \alpha_R J_{SC} = \alpha_F J_{SE} = +a_{12} = +a_{21} .$$

Substitution into (28) yields

$$(32)\qquad \begin{bmatrix} J_E \\ \\ J_C \end{bmatrix} = \begin{bmatrix} J_{SE} & -\alpha_R\, J_{SC} \\ \\ \alpha_F\, J_{SE} & -J_{SC} \end{bmatrix} \begin{bmatrix} \varepsilon^{\frac{qV_{BE}}{kT}} - 1 \\ \\ \varepsilon^{\frac{qV_{BC}}{kT}} - 1 \end{bmatrix}$$

The equivalent circuit for this equation becomes



$\alpha_R J_{CD}$
$\alpha_F J_{ED}$
e
c
J_E
J_C
$J_{ED} = J_{SE}\left(\epsilon^{\frac{qV_{BE}}{kT}} - 1\right)$
$J_{CD} = J_{SC}\left(\epsilon^{\frac{qV_{BC}}{kT}} - 1\right)$
b
J_B

The base current density is simply:

$$J_B = J_E - J_C$$

$$(33)\quad J_B = (1-\alpha_F)J_{SE}\left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right) + (1-\alpha_R)J_{SC}\left(\varepsilon^{\frac{qV_{BC}}{kT}} - 1\right)$$

If we bias the transistor into the active mode: $V_{BE} > 0$, $V_{BC} < < 0$ equation (32) reduces to

$$J_E \simeq J_{SE}\left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right) + \alpha_R\, J_{SC}$$

$$J_C \simeq \alpha_F\, J_{SE}\left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right) + J_{SC}$$

or

$$J_C \simeq \alpha_F\, J_E + (1-\alpha_F\alpha_R)J_{SC}$$

$$\approx \alpha_F\, J_E\ .$$

If α_F is close to unity the base current will be very small, which results in a useful device.

α_F-Forward Current Transport Factor

The question which concerns us next is: How do we get α_F close to unity? According to Eq. (29)

$$(29) \qquad \alpha_F = \frac{a_{21}}{J_{SE}}$$

which, according to Eq. (27) becomes

$$(34) \qquad \alpha_F = \frac{\frac{qD_n}{L_N} n_{PO} \operatorname{csch} \frac{W_B - d_{PE} - d_{PC}}{L_N}}{\frac{qD_{PE}}{L_{PE}} P_{EO} \coth \frac{W_E - d_{ne}}{L_{PE}} + \frac{qD_n}{L_N} \coth \frac{W_B - d_{PE} - d_{PC}}{L_N} n_{PO}}$$

we may re-arrange this expression to obtain

$$(35) \qquad \alpha_F = \frac{\operatorname{sech} \frac{W_B - d_{PE} - d_{PC}}{L_N}}{1 + \frac{D_{pe}}{D_n} \frac{L_N}{L_{pe}} \frac{P_{EO}}{n_{PO}} \frac{\tanh \frac{W_B - d_{PE} - d_{PC}}{L_N}}{\tanh \frac{W_E - d_{ne}}{L_{PE}}}}$$

In order to understand this equation refer to (6) and (7). The minority carrier current which is due to emitter injection is

$$(36) \qquad J_N(x'=0)\Big|_{\text{Due to emitter}} = -\frac{qD_n}{L_N} \coth \frac{W_B - d_{PE} - d_{PC}}{L_N}$$

$$(37) \qquad J_N(x'=w_B)\Big|_{\text{Due to emitter}} = -\frac{qD_n}{L_N} \operatorname{csch} \frac{W_B - d_{PE} - d_{PC}}{L_N} \Delta n(0)$$

The ratio of these two currents is the transport factor

$$(38) \qquad \beta_F^{\#} \equiv \frac{J_N(x'=w_B)}{J_N(x'=0)}\Bigg|_{\text{Due to emitter}} = \operatorname{sech} \frac{W_B - d_{PE} - d_{PC}}{L_N}$$

This transport factor is a part of α_F. We should arrange our transistor in such a way that $\beta_F^{\#} \to 1$. This surely means that $W_B << L_N$. If this condition is satisfied we write

$$\cosh x = \frac{1}{2}\,(\varepsilon^{x} + \varepsilon^{-x})$$

$$\approx \frac{1}{2}\,(1+1) + \frac{1}{2}(1-1)\,x + \frac{1}{2}(1+1)\,\frac{x^2}{2} + ..$$

$$\approx 1 + \frac{1}{2}\,x^2 \quad \text{if } x \text{ is small}$$

Hence

$$\operatorname{sech} x = \frac{1}{\cosh x} \approx \frac{1}{1 + \frac{1}{2}x^2} \approx 1 - \frac{1}{2}\,x^2$$

Thus, if $W_B << L_N$

$$(38a) \qquad \beta_F^{\#} \approx 1 - \frac{1}{2}\left(\frac{W_B - d_{PE} - d_{PC}}{L_N}\right)^2$$

In order to explain the remaining part of α_F let us consider the expansion for the emitter current. We write, in view of (18),

$$(39) \qquad J_E(x'=0)\Big|_{\text{Due emitter injection}} = \left(\frac{qD_{PE}}{L_{PE}} \coth \frac{w_E}{L_{PE}} + \frac{qD_n}{L_N}\, n_{PO}\, \frac{w_B}{L_N}\right)\left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right)$$

The injection efficiency is defined by

$$(40) \qquad \gamma_E \equiv \left.\frac{J_N(x'=0)}{J_E(x'=0)}\right|_{\text{Due to emitter injection}}$$

$$= \left.\frac{1}{1+\dfrac{J_P(x'=0)}{J_N(x'=0)}}\right|_{\text{Due to emitter injection}}$$

But this becomes simply

$$(41) \qquad \gamma_E = \frac{1}{1 + \dfrac{D_{PE}}{D_n}\dfrac{L_N}{L_{PE}}\dfrac{P_{EO}}{n_{PO}}} \quad \frac{\tanh \dfrac{W_B-d_{PE}-d_{PC}}{L_N}}{\tanh \dfrac{W_E-d_{nE}}{L_{PE}}}$$

This is the second part of α_F. We may write

$$(35a) \qquad \alpha_F = \beta_F^{\#} \ \gamma_E$$

$$= \left.\left(\frac{J_N(x'=0)}{J_E(x'=0)} \cdot \frac{J_N(x'=w_B)}{J_N(x'=0)} \right)\right|_{\substack{\text{Due to emitter}\\ \text{injection}}}$$

Thus α_F is the ratio of the emitter-injection minority carrier current which arrives at the collector-base boundary to the total emitter current. It is obvious that we require γ_E to be close to unity. In order to do this assume:

$$D_{PE} \sim D_n, \quad W_B << L_N, \quad W_E >> L_{PE}$$

This would imply a short base and a long emitter. Then

$$\text{(41a)} \quad \gamma_E \approx \frac{1}{1+\frac{P_{EO}}{L_{PE}}\,\frac{(W_B-d_{PE}-d_{PC})}{n_{PO}}}$$

But $\quad P_{EO} = \frac{n_i^2}{N_{DE}} \qquad q\mu_n N_{DE} = \sigma_E = \frac{1}{P_E}$

Hence $\quad \frac{P_{EO}}{n_{PO}} = \frac{N_A}{N_{DE}} \approx \frac{P_E}{P_B}$

where P_E and P_B are the bulk resistivities of the emitter and base respectively. But then

$$\text{(41b)} \quad \gamma_E \approx \frac{1}{1 + \frac{W_B-d_{PE}-d_{PC}}{L_{PE}}\,\frac{P_E}{P_B}}$$

if we dope the emitter very much more than the base $\frac{P_E}{P_B} << 1$. Then

$$\text{(41c)} \quad \gamma_E \approx 1 - \frac{W_B-d_{PE}-d_{PC}}{L_{PE}}\,\frac{P_E}{P_B}$$

and

$$\text{(35b)} \quad \alpha_F \approx 1 - \frac{1}{2}\left(\frac{W_B-d_{PE}-d_{PC}}{L_N}\right)^2 - \frac{W_B-d_{PE}-d_{PC}}{L_{PE}}\,\frac{P_E}{P_B}$$

This expression applies to a well designed device with uniform doping in the base and emitter. The expression for α_R becomes

$$\text{(35c)} \quad \alpha_R \approx 1 - \frac{1}{2}\left(\frac{W_B-d_{PE}-d_{PC}}{L_N}\right)^2 - \frac{W_B-d_{PE}-d_{PC}}{L_{PC}}\,\frac{P_C}{P_B}$$

Normally $P_C >> P_E$, which implies $\alpha_R < \alpha_F$. This condition: $P_C >> P_E$ is normally the result of break-down voltage considerations.

The Ebers-Moll Equivalent

The Ebers-Moll equivalent is perhaps the most powerful tool which we have to help us understand devices. We will demonstrate this with several examples:

(a) Open Emitter:

Suppose we use the collector-base junction as a diode and leave the emitter open. We ask: What kind of a diode is it and what will V_{BE} be? In order to answer this write

$$0 = J_{SE}\ (\varepsilon^{\frac{qV_{BE}}{kT}} - 1) - \alpha_R\ J_{SC}\ (\varepsilon^{\frac{qV_{BC}}{kT}} - 1)$$

or

$$\varepsilon^{\frac{qV_{BE}}{kT}} - 1 = \alpha_R \frac{J_{SC}}{J_{SE}}\ (\varepsilon^{\frac{qV_{BC}}{kT}} - 1)$$

But

$$J_C = \alpha_F\ J_{SE}\ (\varepsilon^{\frac{qV_{BE}}{kT}} - 1) - J_{SC}(\varepsilon^{\frac{qV_{BC}}{kT}} - 1)$$

$$- J_C = J_{SC}\ (1-\alpha_F\alpha_R)\ (\varepsilon^{\frac{qV_{BC}}{kT}} - 1)$$

The saturation current of this diode is therefore $J_S = J_{SC}(1 - \alpha_F\alpha_R)$. The base-emitter voltage becomes

$$\varepsilon^{\frac{qV_{BE}}{kT}} - 1 = - \alpha_R \frac{J_{SC}}{J_{SE}}\ \frac{J_C}{(1-\alpha_F\alpha_R)}\ \frac{1}{J_{SC}}$$

But

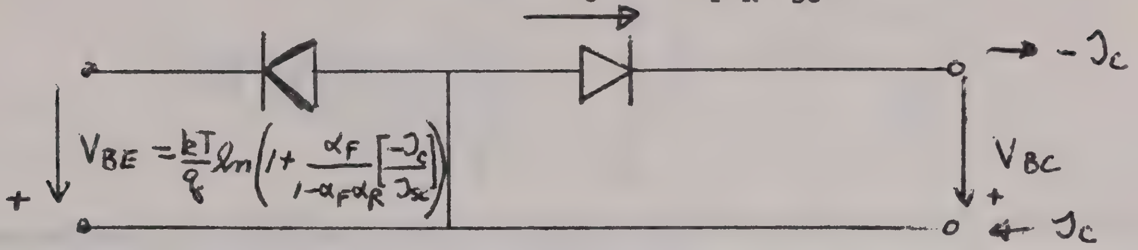
$$\frac{J_{SC}}{J_{SE}} = \frac{\alpha_F}{\alpha_R}$$

Hence

$$V_{BE} = \frac{kT}{q} \ln \left[1 - \frac{\alpha_F}{1-\alpha_F\alpha_R}\ \frac{J_C}{J_{SC}}\right]$$

The device looks as follows:

$$-J_C = (1-\alpha_F\alpha_R)J_{SC}(\varepsilon^{\frac{qV_{BC}}{kT}}-1)$$


$V_{BE} = \frac{kT}{q}\ln\left(1+\frac{\alpha_F}{1-\alpha_F\alpha_R}\left[\frac{-J_c}{J_{sc}}\right]\right)$
+
$-J_c$
V_{BC}
+
J_c

Note that the emitter-base diode is slightly forward biased. Since we know V_{BE} when J_C is known we could calcluate the resulting excess carrier concentration in the base.

If we make V_{BC} negative the collector terminal current becomes

$$J_C \approx (1-\alpha_F\alpha_R)J_{SC} \quad \text{(into collector)}$$

This current is commonly referred to as J_{CBO}. Notice that it appears as part of the collector current in the active mode. (See page 135).

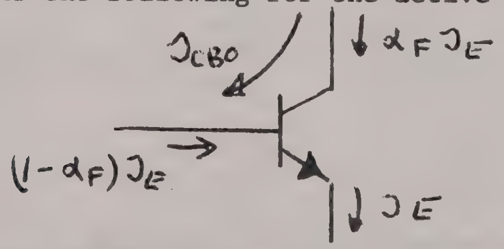
$$J_C \approx \alpha_F J_E + (1-\alpha_F\alpha_R)\, J_{SC}$$

$$= \alpha_F J_E + J_{CBO}$$

Since $J_B = J_E - J_C$

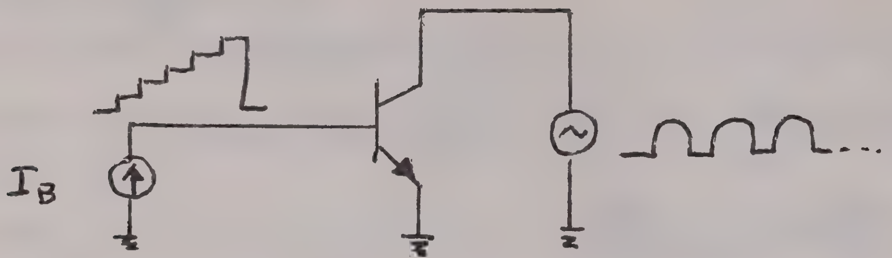
$$= (1-\alpha_F)J_E - J_{CBO}$$

We find the following for the active mode



Grounded Emitter Characteristics:

The most commonly used device characteristics are the I_C versus V_{CE} families in which I_B is used as the parameter. This is accomplished on a curve tracer:



The mathematical description may be derived from the Ebers-Moll equivalent.

$$(1)\quad \begin{bmatrix} I_E \\ I_C \end{bmatrix} = \begin{bmatrix} I_{SE} & -\alpha_R I_{SC} \\ \alpha_F I_{SE} & -I_{SC} \end{bmatrix} \begin{bmatrix} \varepsilon^{\frac{qV_{BE}}{kT}} - 1 \\ \varepsilon^{\frac{qV_{BC}}{kT}} - 1 \end{bmatrix}$$

$$(2)\quad I_B = I_E - I_C$$

$$(3)\quad V_{CE} = V_{BE} - V_{BC}$$

We start by inverting Eq. (1)

$$(4)\quad \begin{bmatrix} \varepsilon^{\frac{qV_{BE}}{kT}} - 1 \\ \varepsilon^{\frac{qV_{BC}}{kT}} - 1 \end{bmatrix} = \frac{1}{I_{SE} I_{SC} (1-\alpha_F \alpha_R)} \begin{bmatrix} I_{SC} & -\alpha_R I_{SC} \\ \alpha_F I_{SE} & -I_{SE} \end{bmatrix} \begin{bmatrix} I_E \\ I_C \end{bmatrix}$$

$$\varepsilon^{\frac{qV_{BE}}{kT}} = 1 + \frac{I_{SC} I_E - \alpha_R I_{SC} I_C}{I_{SE} I_{SC} (1-\alpha_F \alpha_R)}$$

$$\varepsilon^{\frac{qV_{BC}}{kT}} = 1 + \frac{\alpha_F I_{SE} I_E - I_{SE} I_C}{I_{SE} I_{SC} (1-\alpha_F \alpha_R)}$$

The emitter current may be eliminated by using Eq. (2). This yields

$$(5) \qquad \varepsilon^{\frac{qV_{BE}}{kT}} = 1 + \frac{(1-\alpha_R)I_C+I_B}{I_{SE}(1-\alpha_F\alpha_R)}$$

$$(6) \qquad \varepsilon^{\frac{qV_{BC}}{kT}} = 1 \quad \frac{(1-\alpha_F)I_C-\alpha_F I_B}{I_{SC}(1-\alpha_F\alpha_R)}$$

We combine the two equations by using Eq. (3)

$$\left(1 - \frac{(1-\alpha_F)I_C - \alpha_F I_B}{I_{SC}(1-\alpha_F\alpha_R)}\right) \varepsilon^{\frac{qV_{CE}}{kT}} = 1 + \frac{(1-\alpha_R)I_C + I_B}{I_{SE}(1-\alpha_F\alpha_R)}$$

or

$$\varepsilon^{\frac{qV_{CE}}{kT}} - 1 = \left(\frac{1-\alpha_R}{I_{SE}(1-\alpha_F\alpha_R)} - \frac{1-\alpha_F}{I_{SC}(1-\alpha_F\alpha_R)}\varepsilon^{\frac{qV_{CE}}{kT}}\right)I_C$$

$$+ \left(\frac{1}{I_{SE}(1-\alpha_F\alpha_R)} - \frac{\alpha_F\,\varepsilon^{\frac{qV_{CE}}{kT}}}{I_{SC}(1-\alpha_F\alpha_R)}\right)I_B$$

The emitter saturation term may be eliminated by recalling that

$$\alpha_R I_{SC} = \alpha_F I_{SE}$$

$$I_{SE} = \frac{\alpha_R}{\alpha_F} I_{SC}$$

$$I_{SC}(1-\alpha_F\alpha_R)(\varepsilon^{\frac{qV_{CE}}{kT}} -1) = (\frac{\alpha_F}{\alpha_R}(1-\alpha_R) + (1-\alpha_F)\ \varepsilon^{\frac{qV_{CE}}{kT}})I_C$$

$$+ (\frac{\alpha_F}{\alpha_R} - \alpha_F\ \varepsilon^{\frac{qV_{CE}}{kT}})I_B$$

Then

$$(7) \qquad I_C = \frac{\alpha_F(\varepsilon^{\frac{qV_{CE}}{kT}} - \frac{1}{\alpha_R})I_B + (1-\alpha_F\alpha_R)I_{SC}\,(\varepsilon^{\frac{qV_{CE}}{kT}} - 1)}{(1-\alpha_F)\,(\varepsilon^{\frac{qV_{CE}}{kT}} + \frac{\alpha_F}{1-\alpha_F}\,\frac{1-\alpha_R}{\alpha_R})}$$

It is this equation which is displayed on the curve tracer. If $\varepsilon^{\frac{qV_{CE}}{kT}}$ becomes very large the equation reduces to

$$(7a) \qquad I_C = \frac{\alpha_F}{1-\alpha_F} I_B + \frac{1}{1-\alpha_F}(1-\alpha_F\alpha_R)I_{SC}$$

The curves are independent of V_{CE}: They are constant current lines. The separation between successive base current steps may be obtained:

$$I_C(n) = \frac{\alpha_F}{1-\alpha_F}\Delta I_B n + \frac{1}{1-\alpha_F}(1-\alpha_F\alpha_R)I_{SC}$$

$$I_C(n+1) = \frac{\alpha_F}{1-\alpha_F}\Delta I_B(n+1) + \frac{1}{1-\alpha_F}(1-\alpha_F\alpha_R)I_{SC}$$

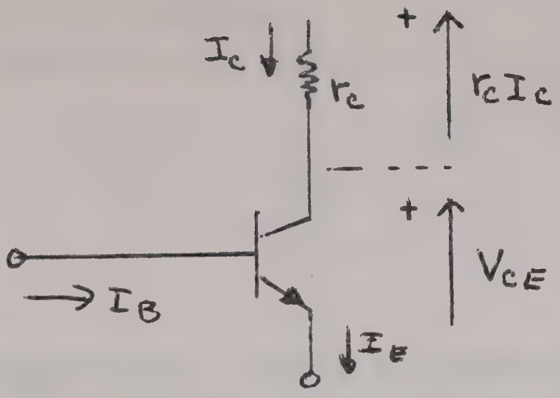
Hence:

$$\Delta I_C = I_C(n+1) - I_C(n) = \frac{\alpha_F}{1-\alpha_F}\Delta I_B$$

or

$$(8)] \qquad \beta_F \equiv \frac{\Delta I_C}{\Delta I_B} = \frac{\alpha_F}{1-\alpha_F}$$

This is the common emitter current amplification factor. The low voltage response of the transistor is complicated. Parasitic elements enter strongly. This is particularily true for the collector resistance. This element, which is not part of the diffusion model, is due to ohmic voltage drop in the collector. We may account for it as follows:



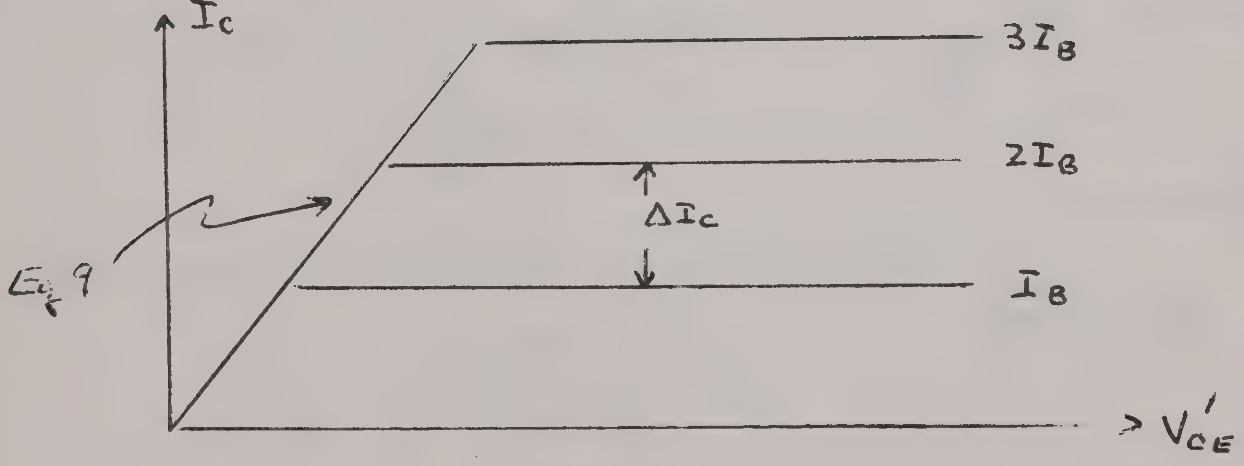
The measured collector to emitter voltage is

$$V'_{CE} = V_{CE} + r_C I_C$$

and Eq. (7) should be modified by letting $V_{CE} = V'_{CE} - r_C I_C$. The true collector to emitter voltage will therefore be zero if

$$V'_{CE} = r_C I_C \tag{9}$$

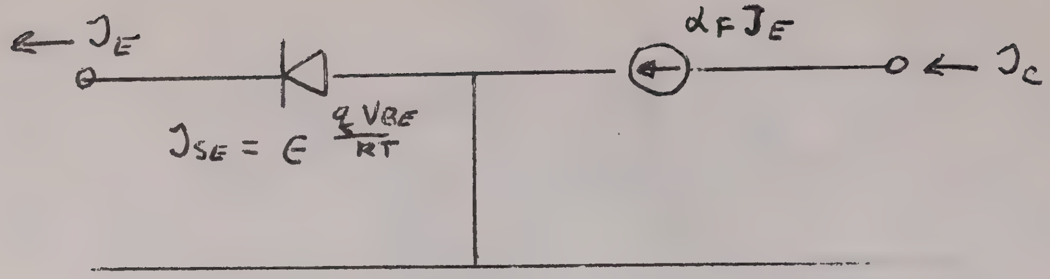
We should therefore obtain a curve as follows:



First Order Common Emitter Characteristics

Transistor Equivalent Circuit

In view of our previous discussion it is evident that the low frequency equivalent for the transistor in the active mode is



This implies that the output impedance is infinite. Unfortunately this is not quite true. In order to understand this consider

$$J_C = \alpha_F\, J_{SE}\, (\varepsilon^{\frac{qV_{BE}}{kT}} - 1) - J_{SC}(\varepsilon^{\frac{qV_{BC}}{kT}} - 1)$$

We define the output admitance by

$$g_C \equiv - \frac{\partial J_C}{\partial V_{BC}} \Bigg|_{\Delta V_{BE}=0} = \frac{\partial \delta_F}{\partial V_{BC}}\, J_{SE}(\varepsilon^{\frac{qV_{BE}}{kT}} - 1) - J_{SC}\, \frac{q}{kT}\, \varepsilon^{\frac{qV_{BC}}{kT}}$$

The last term is very small in the active mode since $V_{BC} \to -\infty$. We will neglect this term. Since

$$J_C \approx \alpha_F\, J_{SE}\, \varepsilon^{\frac{qV_{BE}}{kT}}$$

We may write

$$\frac{\partial J_C}{\partial V_{BC}} \approx \frac{1}{\alpha_F} \frac{\partial \alpha_F}{\partial V_{BC}}\, J_C$$

But

$$\alpha_F \approx 1 - \frac{1}{2} \left(\frac{W_B - d_{PC} - d_{Pe}}{L_N}\right)^2 - \frac{P_E}{P_B}\, \frac{W_B - d_{PC} - d_{Pe}}{L_{PE}}$$

Hence α_F is a function of V_{BC} due to the depletion region d_{PC} which narrows the base. The collector current may be written as

$$J_C \approx \frac{qD_n n_{P0}}{L_N} \frac{1}{\sinh \frac{w_B}{L_N}} \left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right)$$

Hence

$$\frac{\partial J_C}{\partial V_{BC}} = \frac{qD_n n_{P0}}{L_N} \frac{\cosh \frac{w_B}{L_N}}{\sinh^2 \frac{w_B}{L_N}} \frac{1}{L_N} \frac{\partial w_B}{\partial V_{BC}} \left(\varepsilon^{\frac{qV_{BE}}{kT}} - 1\right)$$

$$= J_C \coth \frac{w_B}{L_N} \frac{1}{L_N} \frac{\partial w_B}{\partial V_{BC}}$$

$$= J_C \frac{1}{w_B} \frac{\partial w_B}{\partial V_{BC}} \quad \text{if} \quad \frac{w_B}{L_N} << 1$$

Consider now a linear graded collector-base junction

$$w_B = W_B - d_{PC} - d_{PE} \approx W_B - d_{PC}$$

$$= W_B - \left(\frac{3\varepsilon V_{CB}}{2qa}\right)^{1/3}$$

Here a is the slope of the impurity profile at the junction. Then

$$\frac{\partial w_B}{\partial V_{BC}} = -\frac{1}{3}\left(\frac{3\varepsilon V_{CB}}{2qa}\right)^{1/3} \frac{1}{V_{CB}} = -\frac{d_{PC}}{3V_{CB}}$$

Then

$$\frac{\partial J_C}{\partial V_{BC}} = -\frac{J_C}{3V_{CB}} \frac{d_{PC}}{W_B - d_{PC}}$$

or

$$r_C' = 3 \frac{V_{CB}}{J_C} \left(\frac{W_B}{d_{PC}} - 1\right)$$

Consider now a transistor which has $W_B = .q\mu$, $d_{PC} = .3\mu$ $V_{CB} = 10$ volts, $I_C = 10$ ma. Then

$$r_C = 3 \frac{10}{10x10^{-3}} \left(\frac{.9}{.3} - 1 \right)$$

$$= 6000 \text{ ohms}$$

Surely this is less than ∞. The resistor r_C which shunts the current source should therefore be added.

AC-Diffusion Model

Let us now consider what modifications we must make to account for AC excitation of the device. Consider the minority carrier transport equation in the base:

$$\frac{\partial \Delta n}{\partial t} = - \frac{\Delta n}{Tn} + D_n \frac{\partial^2 \Delta n}{\partial x^2}$$

The laplace transform of this equation is

$$S \Delta N = - \frac{\Delta N}{T_N} + D_n \frac{d^2 \Delta N}{dx^2} \qquad \Delta N = \mathcal{L}_t(\Delta n)$$

We write this as

$$\Delta N - \frac{L_N^2}{1+ST_N} \frac{d^2 \Delta N}{dx^2} = 0$$

If we define the AC-diffusion length by

$$\Lambda_n = \frac{L_N}{\sqrt{1+ST_n}}$$

$$\Delta N - \Lambda_n^2 \frac{d^2 \Delta N}{dx^2} = 0$$

This has solution

$$\Delta N = A\ \varepsilon^{\frac{x}{\Lambda_n}} + B\ \varepsilon^{-\frac{x}{\Delta n}}$$

But this is exactly the form of the DC-solution. Thus, if we assume that we have some excess densities at the base boundaries our solution will take the same form as the DC-solution. Indeed we can get the AC-solution from the DC-solution by the replacement $L_N \to \Lambda_n$.

The boundary conditions are of the form

$$P_n = P_{n0}\ \varepsilon^{\frac{qV}{kT}}$$

If V is time dependent these become

$$P_n = P_{n0}\ \mathcal{L}\left(\varepsilon^{\frac{qV}{kT}}\right)$$

and

$$\Delta P = P_{n0}\ \left(\mathcal{L}\left(\varepsilon^{\frac{qV}{kT}}\right) -1\right)$$

With these ideas in mind consider a long diode. The DC-solution is

$$J = \left(\frac{qD_n}{L_n}\ n_{P0} + \frac{qD_P}{L_P}\ P_{n0}\right)\ \left(\varepsilon^{\frac{qV}{kT}}-1\right)$$

The AC-solution becomes:

$$J = \left(\frac{qD_n n_{P0}}{L_n}\ \sqrt{1+ST_n} + \frac{qD_P P_{n0}}{L_P}\ \sqrt{1+ST_P}\right)\ \left(\mathcal{L}_t\ \left(\varepsilon^{\frac{qV}{kT}}\right) -1\right)$$

In the steady state $S \to j\omega$, where ω is the driving frequency. If we assume that $1 >> \omega T_N$ and $1 >> \omega T_P$ we may expand the square roots:

$$J_S \approx \left[\frac{qD_n n_{P0}}{L_N} + \frac{qD_P P_{n0}}{L_P} + j\omega\left(\frac{qD_n n_{P0} T_n}{L_N \; 2} + \frac{qD_P P_{n0} T_P}{L_P \; 2}\right)\right]$$

The saturation current becomes frequency dependent. The applied voltage is assumed to be

$$V = V_0 + v(t) \qquad\qquad v(t) << V_0$$

Then

$$\varepsilon^{\frac{qV}{kT}} = \varepsilon^{\frac{qV_0}{kT}} \varepsilon^{\frac{qv(t)}{kT}}$$

$$= \varepsilon^{\frac{qV_0}{kT}} \left(1 + \frac{q}{kT} v(t)\right)$$

Hence

$$\mathcal{L}_t\left(\varepsilon^{\frac{qV}{kT}}\right) = \varepsilon^{\frac{qV_0}{kT}} \left(1 + \frac{q}{kT} V(s)\right)$$

in the steady state this becomes

$$\mathcal{L}_t\left(\varepsilon^{\frac{qV}{kT}}\right) = \varepsilon^{\frac{qV_0}{kT}} \left(1 + \frac{q}{kT} v_0 \varepsilon^{j\omega t}\right)$$

But then the diode current is

$$J = \left(\frac{qD_n n_{P0}}{L_N} + \frac{qD_P P_{n0}}{L_P}\right) \varepsilon^{\frac{qV_0}{kT}} + \left(\frac{qD_n n_{P0}}{L_N} \times \left(1 + j\omega \frac{T_n}{2}\right) + \right.$$

$$\left. \frac{qD_P P_{n0}}{L_P} \left(1 + j\omega \frac{T_P}{2}\right)\right) \varepsilon^{\frac{qV_0}{kT}} \frac{q}{kT} v_0 \; \varepsilon^{j\omega t}$$

The first term is the DC-solution; the second term the AC-solution. We define the AC-admittance as

$$Y = \frac{J_{AC}}{v_0 \varepsilon^{j\omega t}} = \left(\frac{qD_n n_{P0}}{L_N} \left(1 + j\omega \frac{T_n}{2} \right) + \frac{qD_P P_{n0}}{L_P} \left(1 + j\omega \frac{T_P}{2} \right) \right) \varepsilon^{\frac{qV_0}{kT}}$$

q/kT

$$= G + j\omega C$$

Where

$$G = \left(\frac{qD_n n_{P0}}{L_n} + \frac{qD_P P_{n0}}{L_P} \right) \varepsilon^{\frac{qV_0}{kT}}$$

$$C = \left(\frac{qD_n n_{P0}}{L_n} \frac{T_n}{2} + \frac{qD_P P_{n0}}{L_P} \frac{T_P}{2} \right) \varepsilon^{\frac{qV_0}{kT}}$$

These four elements: G_p, G_n, C_P, C_N constitute the small signal equivalent for the diode. The network is referred to as the diffusion equivalent.

In a transistor we have two diodes: E-B and BC. These may be represented by a diffusion network.

There is an additional frequency dependence in the transistor. The device is based on the assumption that

$$W_B << L_N$$

This becomes

$$W_B << \left| \frac{L_N}{\sqrt{1 + j\omega T_N}} \right|$$

Obviously this condition cannot be satisfied for high enough frequencies. But this implies that α_F itself is frequency dependent. This is due to the transport factor

$$\beta_F^{\#} = \left[\cosh \left(\frac{W_B - d_{PC} - d_{pe}}{L_N} \sqrt{1 + j\omega T_n} \right) \right]^{-1}$$

By manipulating this equation we can show that

$$\alpha_F = \frac{\alpha_{F0}}{1+j\frac{\omega}{\omega\alpha}}$$

where

$$\omega_\alpha = \frac{1}{T_{N1}} \frac{\pi^2}{4} \left(\frac{L_N}{W_B-d_{PC}-d_{pe}}\right)^2$$

REVIEW OF MODERN PHYSICS

The Basic tools for the analysis and understanding of conduction phenomena in solids are contained in what we might loosely call quantum mechanics. The early name of this segment of modern physics is wave mechanics, which would imply that some understanding of a "wave" is needed.

Consider then the basic ideas which we associate with a wave: space and time. Thus

$$\phi(x,t) = A \sin (kx-\omega t) \tag{1}$$

represents a plane wave. The word "plane" arises as follows. The phase of the wave is defined by

$$P_o = kx-\omega t \tag{2}$$

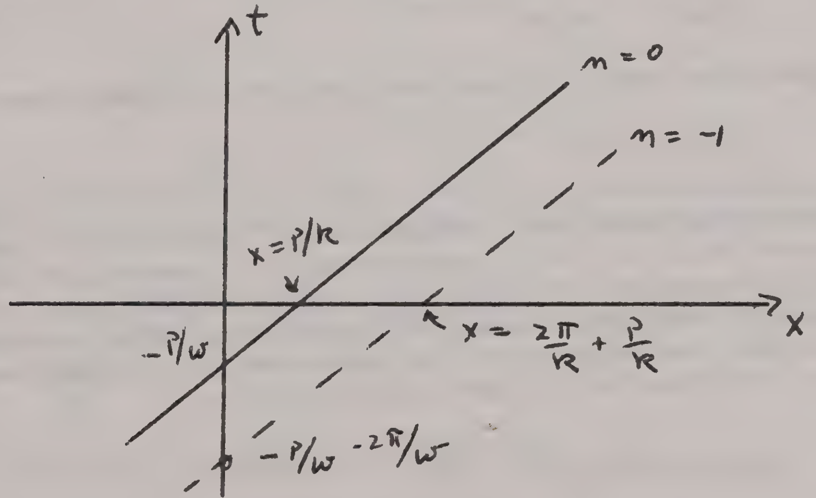
But $\sin\theta = \sin(\theta+2\pi n)$. Thus, more accurately

$$P = kx-\omega t + 2\pi n \tag{3}$$

For points of constant phase P is a constant. Here

$$t = \frac{k}{\omega} x + \frac{2\pi n}{\omega} - \frac{P}{\omega} \tag{4}$$

Represents a family of straight lines on the x-t-plane.



Note that the distance , measured along the x-axis, between neighboring equi-phase planes is

$$\Delta x = \frac{P}{k} + \frac{2\pi}{k} - \frac{P}{k} \tag{5}$$

$$\Delta x = \frac{2\pi}{k}$$

but this should be exactly one wavelength. Hence

(5a) $$\lambda = \frac{2\pi}{k}$$

The velocity with which the constant phase points move is readily obtained by differentiating (4)

(6) $$1 = \frac{k}{\omega}\ \frac{dx}{dt}$$

$$\frac{dx}{dt} = \frac{\omega}{k} = \frac{\omega}{2\pi}\ \lambda = f\lambda = v_\phi$$

This velocity is called the phase velocity of the wave. The parameter k is the wave number. The fact that the constant phase contours would form planes in a three dimension space x,y,z yields the name "plane" wave.

We may write equation (1) somewhat differently.

$$\begin{aligned}\phi(x,t) &= A \sin(kx-\omega t)\\ &= A \sin(\frac{2\pi}{\lambda}\ x-2\pi ft)\\ &= A \sin 2\pi(\frac{x}{\lambda} - ft)\\ &= A \sin \frac{2\pi}{\lambda}(x-v_\phi t)\\ \phi(x,t) &= \text{Im}\ (A\varepsilon^{j(kx-\omega t)})\end{aligned}$$

or, if it is understood that the imaginary part will be taken even

(7) $$\psi(x,t) = A\varepsilon^{j(kx-\omega t)}$$

Suppose we examine this plane wave somewhat more. Equation (7) suggests that it extends from $x = -\infty$ to $x = +\infty$. The location of a particular point on the sine-wave might be specified by starting the phase at that point. Note, however, that there are some conceptional difficulties in this location idea. In order to clarify this let us consider the superposition of the plane waves. The first one is specified by (7), the second by (8):

(8) $$\psi_2(x,t) = A_2\ \varepsilon^{j(k+\Delta k)x\ -\ (\omega+\Delta\omega)t)}$$

This second wave has a slightly different frequency. It therefore has a slightly different wavelength and wave number. The phase velocity of the two waves may be the same. If this is so the specification of $\Delta\omega$ will also specify Δk:

(9) $$\frac{\omega+\Delta\omega}{k+\Delta k} = \frac{\omega}{k}$$

$$\Delta k = \frac{k}{\omega}\ \ \Delta\omega$$

the total wave is the sum of the two components

$$(10) \qquad \psi_T = \psi + \psi_2$$

$$= A\varepsilon^{j(kx-\omega t)} \left[1 + \frac{A_2}{A} \varepsilon^{j(\Delta kx-\Delta\omega t)}\right]$$

But this is nothing more than an amplitude modulated wave. Its maximum is

$$\psi_T/_{max} = (A + A_2)\ \varepsilon^{j(kx-\omega t)}$$

its minimum is

$$\psi_T/_{min} = (A-A_2)\ \varepsilon^{j(kx-\omega t)}$$

The maximum occurs whenever

$$(11) \qquad \Delta kx-\Delta\omega t = 2\pi m \qquad m = 0,\pm1,\pm2....$$

The velocity of the maximum is

$$(12) \qquad \frac{dx}{dt} = \frac{\Delta\omega}{\Delta k}$$

or, in the unit

$$(12a) \qquad v_g = \frac{d\omega}{dk}$$

This quantity is defined as the group velocity of the total wave. Note that if the phase velocity is a constant, i.e. independent of frequency; then

$$\omega = kv_\phi$$

$$\frac{d\omega}{dk} = v_\phi$$

if, however, the phase velocity is a function of frequency then

$$(13) \qquad \frac{d\omega}{dk} = v_\phi + k\,\frac{dv_\phi}{d\omega}\,\frac{d\omega}{dk}$$

We may summarize our ideas on plane waves by working an example. Consider a very long resistor which is shunt-loaded by capacitors. This network is in the limit an RC-transmission line. Its analysis proceeds as follows:

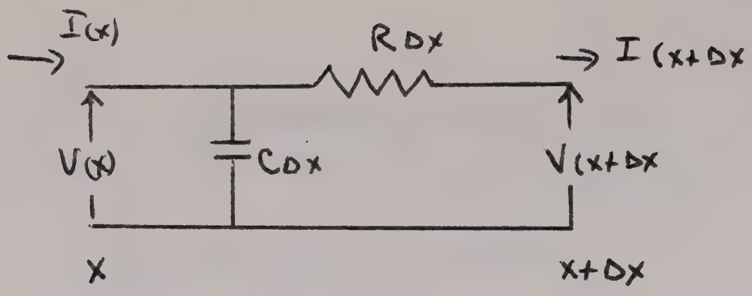


Fig. 1

Fundamental RC-Line Section

in order to write the differential equations for this network consider the loop equation

$$V(x) = R\Delta x\ (I(x+\Delta x)) + V(x+\Delta x)$$

$$V(x+\Delta x) = V(x) + \Delta V$$

$$= V(x) + \frac{\partial V}{\partial x}\Delta x$$

here

$$\frac{\partial V}{\partial x} = -\ RI(x+\Delta x)$$

the node equation at x yields

$$I(x) = I(x+\Delta x) + C\Delta x\,\frac{\partial V}{\partial t}$$

Hence $$\frac{\partial I}{\partial x} = -\ C\,\frac{\partial V}{\partial t}$$

if we take the limit case $\Delta x \to 0$ this system becomes

(14) $$\frac{\partial V}{\partial x} = -\ RI$$

(15) $$\frac{\partial I}{\partial x} = -\ C\,\frac{\partial V}{\partial t}$$

by differentiating one obtains

(16) $$\frac{\partial^2 V}{\partial x^{\cancel{3}2}} = -\ R\,\frac{\partial I}{\partial x}$$

$$\frac{\partial^2 V}{\partial x^{\cancel{3}2}} = RC\,\frac{\partial V}{\partial t}$$

if it is assumed that the solution to (16) has the form

$$V(x,t) = \hat{V}(x)\varepsilon^{j\omega t}$$

Equation (16) becomes

(16a) $$\frac{d^2\hat{V}}{dx^2} = j\omega RC\hat{V}$$

This equation permits the solution

(17) $$V(x,t) = V_o\varepsilon^{-\sqrt{j\omega RC}\,x + j\omega t}$$

where V_o is some constant. We may write this equation somewhat differently by noting that $\sqrt{j} = (\varepsilon^{j\frac{\pi}{2}})^{\frac{1}{2}} = \frac{1+j}{\sqrt{2}}$. Then

(18) $$V(x,t) = V_o\varepsilon^{-\sqrt{\frac{\omega RC}{2}}\,x}\;\varepsilon^{j(\omega t - (\sqrt{\frac{\omega RC}{2}})x)}$$

The phase on this wave is given by

$$P = \omega t - \sqrt{\frac{\omega RC}{2}}\,x$$

The comparison of this expression with Eq. (3) yields

$$k = \sqrt{\frac{\omega RC}{2}} \quad \text{or} \quad \frac{2k^2}{RC} = \omega$$

Hence $$v_G = \frac{d\omega}{dk} \qquad v_\phi = \frac{\omega}{k}$$

$$v_\phi = \frac{2k}{RC} = \sqrt{\frac{2\omega}{RC}}$$

$$v_G = \frac{4k}{RC} = 2\sqrt{\frac{2\omega}{RC}}$$

Phase velocity and group velocity differ.
In summary then:

(1) A wave is described by space and time coordinates.

(2) The wavenumber is $k = \frac{2\pi}{\lambda}$

(3) The phase velocity is $v_\phi = \frac{\omega}{k}$

(4) $v_\phi = f\lambda$

(5) The wave group velocity is $v_G = \frac{d\omega}{dk}$

(6) The existence of a group velocity implies the existence of more than one frequency in the wave. It leads to the concept of the wave packet.

WAVE PACKET

The equation for a plane wave is given by (7).

$$\psi = A_1 \varepsilon^{j(kx-\omega t)}$$

We may write this as

$$\psi = u(k,x)\ \varepsilon^{-j\omega t}$$

where (19) $u(k,x) = A_1 \varepsilon^{jkx}$
Is the so called amplitude function.

Suppose now that we assume that we have a continuum of equal amplitude functions each with a slightly different value of k. Then the total amplitude function will be given by

$$(20) \quad u_T(x) = \int_{k_0-\Delta k}^{k_0+\Delta k} \varepsilon^{jkx} dk$$

Here we assume that equal amplitude functions of height unity exist for

$$k_0-\Delta k \leq k \leq k_0+\Delta k$$

or

$$\Delta k \leq k-k_0 \leq \Delta k$$

equation (20) may be integrated to yield

$$(20a) \qquad u_T(x) = 2\Delta k \ \frac{\sin\Delta kx}{\Delta kx} \ \varepsilon^{jk_0x}$$

Note that if we write

$$(21) \qquad f(k) = u(k-(k_0-\Delta k)) - u(k-(k_0+\Delta k))$$

then equation (20) can be written as

$$(20b) \qquad u_T(x) = \int_{-\infty}^{+\infty} f(k)\varepsilon^{jkx} dk$$

therefore, the amplitude function $u_T(x)$ is the Fourier transform of f(k), and k and x are Fourier variables.

The relationship between $u_T(x)$, the amplitude function, and $\psi_T(x,t)$, the wave function is not so obvious. We write

$$(22) \qquad \psi_T(x,t) = \int_{-\infty}^{+\infty} f(k)\varepsilon^{j(kx-\omega t)}dk$$

$$= \int_{-\infty}^{+\infty} f(k)\varepsilon^{jk(x-v_\phi t)}dk$$

$$= \int_{-\infty}^{+\infty} (f(k)\varepsilon^{-jkv_\phi t})\varepsilon^{jkx}dk$$

Suppose now that v_ϕ is a constant. We may then use the following theorem from Fourier transforms:

$$\text{if} \qquad u_T(x) = \int_{-\infty}^{+\infty} f(k)\varepsilon^{jkx}dk = F(f(k))$$

$$\text{then} \qquad u_T(x-a) = \int_{-\infty}^{+\infty} (f(k)\varepsilon^{-jka})\varepsilon^{jkx}dx$$

$$= F(f(k)\varepsilon^{-jka})$$

Where a is a positive real constant.
This theorem or a detailed integration allows us to write

$$(22a) \qquad \psi_T(x,t) = 2\Delta k \, \frac{\sin\Delta k(x-v_\phi t)}{\Delta k(x-v_\phi t)} \, \varepsilon^{jk_0(x-v_\phi t)}$$

therefore if v_ϕ is a constant the relationship between the amplitude and wave-function becomes

$$u_T(x) = \psi_T(x,t=0)$$

Note that Eq. (22a) is again an amplitude modulated sinusoidal wave. The wave has a single maximum which is given by

$$0 = \Delta k(x-v_\phi t)$$

$$0 = \Delta kx - \Delta\omega t$$

This maximum travels with a group velocity which is equal to the phase velocity. This condition is due to the assumption that the phase velocity is a constant. The single rather than multiple maximum which we had earlier is due to the use of the Fourier integral rather than the Fourier series. There is one more consideration which we might find of interest later on. Suppose we ask the question: under what condition will f(k) represent a wave packet?

In order to answer this lets study the amplitude function

$$u_T(x) = \int_{-\infty}^{+\infty} f(k)\varepsilon^{jkx}dk$$

note that if

$$|kx| << 1$$

$$u_T(x) \neq 0 \text{ if } f(k) \neq 0$$

in order words the intergral will have a non-zero value as long as

(23) $$|\Delta k\Delta x| << 1$$

Here Δk and Δx refer to suitable intervals in k-space and x-space respectively. If we now state that $f(k) \approx 0$ outside this k interval then $u_T(x)$ will be non-zero only over the interval Δx. Hence $u_T(x)$ will be the amplitude function associated with a wave packet.

PRACTICLE CONCEPTS

In order to understand what is meant by a particle let us start with Einstein's Mass-energy equivalence. It's origin resides in the fact that the mass of a nucleus is slightly less than the sum of the masses of its constituents.

(1) $$E_T = mc^2$$

The mass in this equation is the relativistic mass

(2) $$m = \frac{m_0}{[1-(\frac{v}{c})^2]^{1/2}}$$

Here $m_0 = m(v=0)$ is the rest mass of the particle. The particle momenum is

(3) $$\bar{P} = m\bar{v}$$

Lets consider the energy in some detail. If $v = 0$

(4) $$E_T(v=0) = E_{REST} = m_0c^2$$

It appears as if E_T is the total energy. It contains terms which depend on velocity and some form of rest energy, which we would normally associate with potential energy. Let us explore this a little bit further. That part of the energy which is associated with motion is

$$E_k = E_T - E_{REST}$$

(5)

$$E_k = m_0c^2 \left[\frac{1}{\sqrt{1-(\frac{v}{c})^2}} - 1\right]$$

consider now the unit as $\frac{v}{c} << 1$.

$$\frac{1}{\sqrt{1-(\frac{v}{c})^2}} \approx \frac{1}{\sqrt{1-(\frac{v}{c})^2+(\frac{1}{2}(\frac{v}{c})^2)^2}} = \frac{1}{1-\frac{1}{2}(\frac{v}{c})^2}$$

$$\approx 1 + \frac{1}{2}(\frac{v}{c})^2$$

Hence

$$E_k \approx m_0c^2[1 + \frac{1}{2}(\frac{v}{c})^2 - 1]$$

(5a)

$$E_k \approx \frac{1}{2}\, m_0v^2$$

This is the familiar classical expression for kinetic energy. Note that it refers to a particle with constant mass. The total energy of this non-relativistic particle is

(1a) $$E_T = \frac{1}{2}\, m_0v^2 + m_0c^2$$

The rest energy appears to be a potential energy term. Note, however, that we are essentially talking about a free particle. In electromagnetic theory we relate field and potential by requiring the field to be the negative spatial gradient or the potential. Field in turn is associated with force. A constant potential will not result in a force on the particle. Hence we speak of a free particle: one on which no forces act. The argument is exactly the same here. We may conclude that Einstein is talking about a relativistic, free particle with the property

(I) (1) $E_T = mc^2$ (1a) $E_{REST} = m_0c^2$

(II) (2) $\bar{P} = m\bar{v}$ (1b) $m = m_0[1-(\frac{v}{c})^2]^{-\frac{1}{2}}$

(III) (6) $pc = [E_T^2-E_{REST}^2]^{1/2}$

The last equation is a consequence of (1), (2) and (1b).

THE PHOTON

Consider a photon. It has a frequency γ and wavelength λ. It is a wave and travels with the speed of light. Its energy, according to Planck is

(1) $$E_T = h\gamma$$

and, since it is a wave

(2) $$v_\phi = \lambda\gamma$$

There are some difficulties, however. Planck states that the energy of the photon cannot be subdivided. In the compton effect the photon acts as a particle. If it is a particle what is its rest mass? Can it be both wave and particle? If it can how are the two properties energy and momentum conserved?

In order to study these questions let us start with the condition

(3) $$E_{PART} = E_{WAVE}$$

or $$mc^2 = h\gamma$$

or (4) $$m = \frac{h\gamma}{c^2}$$

Thus, if we assign the mass $m = \frac{h\gamma}{c^2}$ to the photon - (particle) it will have the same energy as the photon - (wave).

Momentum is

$$P = mv$$

$$= (\frac{h\gamma}{c^2})v$$

Here v is the particle velocity. But the photon-particle travels with the speed of light. Hence

(5) $$P = \frac{h\gamma}{c}$$

if we now assume that the phase velocity of the wave is also equal to the speed of light we obtain from (2) and (5)

(5a) $$P = \hbar k$$

for the momentum of the photon-wave. The remaining particle equation

(6) $$(pc)^2 = E_T^2 - E_{REST}^2$$

becomes

$$(h\gamma)^2 = (h\gamma)^2 - E_{REST}^2$$

From which we conclude that the rest energy and therefore the rest mass of the particle is zero. An alternate expression for Eq. (6) is

$$\text{(6a)} \qquad \omega = ck$$

here

$$v_\phi = \frac{\omega}{k} = c$$

$$v_G = \frac{d\omega}{dk} = c$$

the transformation from a particle to a wave is successful. The wave description is

$$\text{I.} \quad E_T = h\gamma \qquad (E_T = mc^2)$$

$$\text{II.} \quad P = \hbar k \qquad (P = mv)$$

$$\text{III.} \quad \omega = q(k) \qquad ((pc)^2 = E_T^2 - E_{REST}^2)$$

WAVE-PARTICLE DUALITY

In the preceeding section we studied the photon as a particle. Let us now examine an electron as a wave. In order to do this we start with a relativistic, free electron.

$$\text{(1)} \qquad E_{PART} = E_{WAVE}$$

$$mc^2 = h\gamma$$

$$\text{(2)} \qquad P_{PART} = P_{WAVE}$$

$$mv_p = \hbar k$$

combining the two equations yields:

$$v_p = \frac{\hbar k}{m} = \frac{k}{2\pi}\frac{c^2}{\gamma} = \frac{c^2}{\lambda\gamma}$$

But

$$v_\phi = \lambda\gamma$$

Here

$$\text{(3)} \qquad v_p v_\phi = c^2$$

Thus, in order for both points of view to be valid we must have the property that the particle velocity times the phase velocity of the wave is equal to the speed of light squared. This still doesn't tell us how to transform a particle of velocity v_p to a wave. However, we have one more equation:

$$(pc)^2 = E_T^2 - E_{REST}^2 \tag{4}$$

or

$$(\hbar kc)^2 = (\hbar\omega)^2 - (m_0c^2)^2$$

$$\omega = kc\left[1 + \left(\frac{m_0c}{\hbar k}\right)^2\right]^{1/2} \tag{4a}$$

The quantity

$$\lambda_{ct} = \frac{h}{m_0c}$$

is commonly referred to as the compton wavelength of a particle. The associated wavenumber is

$$k_{ct} = \frac{2\pi}{\lambda_{cT}} = \frac{m_0c}{\hbar}$$

then

$$\omega = kc\left[1 + \left(\frac{k_c}{k}\right)^2\right]^{1/2} \tag{4b}$$

We call this relationship the dispersion relationship. Since the phase velocity is defined by $v_\phi = \frac{\omega}{k}$ it becomes

$$v_\phi = c\left[1 + \left(\frac{k_c}{k}\right)^2\right]^{1/2} \tag{5}$$

The group velocity becomes $v_G = \frac{d\omega}{dk}$

$$v_g = \frac{c}{\left[1+\left(\frac{k_c}{k}\right)^2\right]^{1/2}} \tag{6}$$

Or

$$v_Gv_\phi = c^2 \tag{7}$$

But in view of Eq. (3) this results in the immediate conclusion that group velocity and particle velocity are the same thing. A little more accurately: the particle velocity becomes the group velocity in the wave picture. We emphasize this by writing

$$(8) \qquad v_p = v_G = \frac{d\omega}{dk} = \frac{1}{\hbar}\,\frac{dE_T}{dk}$$

Suppose we ask next: What modifications can be made in the wave picture if the particle becomes non-relativistic? Since this means $\frac{v_p}{c} \ll 1$ we may use (6) to write

$$\frac{v_G}{c} = \frac{1}{[1+(\frac{k_c}{k})^2]^{1/2}} \ll 1$$

or

$$k_c \gg k$$

applying this approximation to (4b) yields

$$(9) \qquad \begin{aligned} \omega &= k_c c \;\; [1 + (\frac{k}{k_c})^2]^{1/2} \\ \omega &\approx k_c c \;\; [1 + \frac{1}{2}\,(\frac{k}{k_c})^2] \end{aligned}$$

This is the non-relativistic dispersion relationship. Note that

$$(9a) \qquad \frac{d\omega}{dk} = \frac{\hbar k}{m_0} \quad \text{or} \quad \frac{1}{h}\,\frac{dE_T}{dk} = v_p$$

and

$$(9b) \qquad \frac{d^2\omega}{dk^2} = \frac{\hbar}{m_0} \quad \text{or} \quad \frac{1}{h^2}\,\frac{d^2E_T}{dk^2} = \frac{1}{m_0}$$

These two equations have very important consequences. The newtonian force equation

$$F = \frac{d}{dt}\,P$$

becomes

$$F = \hbar\,\frac{dk}{dt}$$

in its alternate form:

$$F = m_0\,\frac{dv_p}{dt}$$

it becomes

$$F = \left(\frac{1}{\frac{1}{\hbar^2}\,\frac{d^2E_T}{dk^2}}\right)\left(\frac{1}{\hbar}\,\frac{d^2E_T}{dk^2}\,\frac{dk}{dt}\right)$$

Wave	Particle
(1) $E_T = h\gamma$	(1) $E_T = mc^2$
(2) $p = \hbar k$	(2) $p = mv_p$
(3) $E_{REST} = \hbar k_c c$	(3) $E_{REST} = m_0 c^2$
(4) $v_p = \frac{d\omega}{dk} = \frac{1}{\hbar}\frac{dE_T}{dk}$	(4) v_p
(5) $E_T(k) = \hbar kc\ [1+(\frac{k_c}{k})^2]^{1/2}$	(5) $(pc)^2 = E_T^2 - E_{REST}^2$
(6) $v_\phi = \frac{\omega}{k}$	(6) $v_\phi = \frac{E_T}{p} = \frac{c^2}{v_p}$

Notation: $\hbar = \frac{h}{2\pi}$

$$k_c = \frac{m_0 c}{\hbar}$$

Wave Formulas:

$$k = \frac{2\pi}{\lambda}$$

$$\omega = 2\pi\gamma$$

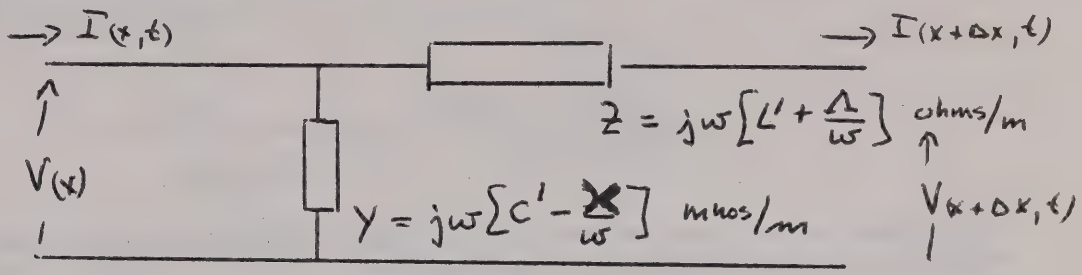
$$v_\phi = \lambda\gamma$$

Special Case: $v << c$

(7) $E_T = \hbar k_c c\,(1 + \frac{1}{2}(\frac{k}{k_c})^2)$	(7) $E_T = m_0 c^2 + \frac{1}{2} m_0 v_p^2$
(8) $m_0 = \frac{1}{\hbar-2\frac{d^2E_T}{dk^2}}$	(8) m_0

THE DISPERSION RELATIONSHIP

In order to illustrate the meaning of the dispersion relationship let us consider a transmission line problem.



The fundamental structure of such a line is shown above. Note that we have frequency dependent L and C elements. L,C,Λ and κ are constants. Harmonic time dependence is assumed. Then

(1) $$\frac{\partial v}{\partial x} = -j\omega LI$$

(2) $$\frac{\partial I}{\partial x} = -j\omega CV$$

or

(3) $$\frac{\partial^2 V}{\partial x^2} + \omega^2 LCV = 0$$

We now assume that the voltage along the line has the solution

$$V = A\varepsilon^{-\gamma x}$$

where

$$\begin{aligned}\gamma^2 &= -\omega^2 LC\\ &= -(\omega L'+\Lambda)(\omega C'-\kappa)\\ &= -[\omega^2 L'C'+\omega(C'\Lambda-L'\kappa)-\kappa\Lambda]\end{aligned}$$

Suppose we pick our line parameters so that

$$L'C' = \frac{1}{c^2}, \quad \frac{L'}{C'} = \frac{\Lambda}{\kappa}, \quad \kappa\Lambda = k_c^2$$

Then $$\gamma = \pm j\left(\left(\frac{\omega}{c}\right)^2 - k_c^2\right)^{1/2}$$

The voltage along the transmission line becomes

$$(1) \qquad \hat{V}(x,t) = A\varepsilon^{j[((\frac{\omega}{c})^2-k_c^2)^{1/2}\,x-\omega t]}$$

but this implies that

$$k^2 = (\frac{\omega}{c})^2-k_c^2$$

$$\omega = ck_c[1+(\frac{k}{k_c})^2]^{1/2}$$

our transmission line has exactly the same dispersion relationship as our free, relativistic particle in the wave picture.

Examination of (1) yields the conclusion that $\hat{V}(x,t)$ is also a solution to the following equation

$$(2) \qquad -\hbar^2 \frac{\partial^2}{\partial t^2}\hat{V}(x,t) = -\hbar^2c^2 \frac{\partial^2}{\partial x^2}\hat{V}(x,t) + (m_0c^2)^2\,\hat{V}(x,t)$$

This equation is Schroediger's equation for a relativistic, free particle.

Suppose now that we want to study the behavior of a wave packet on the transmission line. Note that (1) is the solution for the voltage for harmonic time dependence. But any ω will work and the equation is linear. Hence

$$\phi(x,t) = \int_{-\infty}^{+\infty} f(k)\varepsilon^{j(kx-\omega t)}\,dk$$

will also be a solution. For $f(k)$ we already know the desired character in order to get a wave packet. A suitable function is

$$(3) \qquad f(k) = \varepsilon^{-\frac{k^2}{2\Delta k^2}}$$

Here Δk is a neasure of the width of the packet in k-space. In order to study the effect of the dispersion relationship let us write

$$\omega(k) = \omega(k_0)+\frac{d\omega}{dk}\Big|_{k_0}(k-k_0)+\frac{1}{2}\frac{d^2\omega}{dk^2}\Big|_{k_0}(k-k_0)^2+\ldots$$

or

$$(4) \qquad \omega(k) = \omega_0+\beta(k-k_0)+\frac{1}{2}\alpha(k-k_0)^2$$

The resulting wavefunction is

(5) $$\phi(x,t) = \varepsilon^{j(k_0x-\omega_0t)} \int_{-\infty}^{+\infty} f(k)\,\varepsilon^{j[(k-k_0)(x-\beta t)-\frac{\alpha}{2}(k-k_0)^2t]}\,dk$$

this equation may be integrated. The solution is

(6) $$\phi^*\phi = \frac{2\pi\Delta k^2}{[1+(\alpha t\Delta k^2)^2]^{1/2}}\;\varepsilon^{-\Delta k^2\frac{(x-\beta t)^2}{1+(\alpha t\Delta k^2)^2}}$$

We are interested in the width of the packet as a function of time. Let us estimate this by setting

$$\frac{1}{\varepsilon} = \varepsilon^{-\Delta k^2\frac{(x-\beta t)^2}{1+(\alpha t\Delta k^2)^2}}$$

or

(7) $$(x-\beta t) = \frac{1}{\Delta k}\,[1+(\alpha t\Delta k^2)^2]^{1/2}$$

$$= \Delta x(t)$$

Note that the width of the packet at $t = 0$ is given by $\frac{1}{\Delta k}$. It increases as a function of time. In particular,

$$\Delta x(t)\ \Delta k \geq 1$$

or (8) $$\Delta x\ \Delta p > \hbar$$

This equation is the uncertainty principle. It simply states that it is impossible to have exact information on location and momentum of a particle simultaneously. The idea of a precise value is replaced by a range of values in which the particle exists. The idea of probability occurs.

The uncertainty principle extends to all Fourier transfomr pairs. Thus

(9) $$\Delta t\ \ \Delta\omega \geq 1$$

$$\Delta t\ \ \Delta E \geq \hbar$$

NON-RELATIVISTIC PARTICLE

It was shown earlier that the dispersion relationship for the non-relativistic free particle is of the form

$$(1) \qquad \omega = ck_c + \frac{1}{2}\frac{ck^2}{k_c}$$

or

$$(1a) \qquad E_T = m_o c^2 + \frac{1}{2}\frac{\hbar^2 k^2}{m_o}$$

Note that the first term of (1a) may be considered as the potential energy of the particle:

$$(2) \qquad V_o = m_o c^2$$

since

$$(3) \qquad k = \left[\frac{2m_o}{\hbar^2}(\hbar\omega - m_o c^2)\right]^{\frac{1}{2}}$$

The solution to the transmission line voltage becomes

$$(4) \qquad \hat{V}(x,t) = A\varepsilon^{j[(\frac{2m_o}{\hbar^2}(\hbar\omega - m_o c^2))^{\frac{1}{2}} x - \omega t]}$$

$$= A\varepsilon^{j[(\frac{2m_o}{\hbar^2}(E_T - V_o))^{\frac{1}{2}} x - \frac{E_T}{\hbar} t]}$$

Examination of this equation yields the conclusion that $\hat{V}(x,t)$ is a solution of the equation

$$(5) \qquad -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V_o \quad \hat{V}(x,t) = j\hbar\frac{\partial}{\partial t}\hat{V}(x,t)$$

This is Schroedinger's equation for a particle in a constant potential. If we assume that

$$\hat{V}(x,t) = \psi(x)\varepsilon^{-j\frac{E_T}{\hbar}t}$$

which is essentially the steady state assumption from circuit theory the equation becomes

$$(6) \qquad (-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V_o) \psi(x) = E_T \psi(x)$$

This equation is generalized by removing the restriction that V_o is a constant. It is then called Schroedinger's time independent equation.

The question becomes now: What can this equation tell us? It would appear that we should start the answer to this question by examining the physical significance of ψ. In order to do this we note first of all that ψ is a complex quantity. But then $\psi^*\psi$ is a real quantity. $\psi^*\psi$ is most likely a function of x. Note that we can say nothing about the dimension of ψ. Equation (6) may be multiplied by any constant but $\psi^*\psi$ appears to be some sort of intensity which is defined at the point x. In view of the wave-particle picture this intensity would be related to the location problem of the particle in the wave picture. In that sense it makes sense to call $\psi^*\psi\, dx$ the probability of finding the particle in the interval $[x+dx,x]$. The probability of finding the particle somewhere in one dimensional space is unity. Hence

$$(7) \qquad 1 = \int_{-\infty}^{+\infty} \psi^*\psi \, dx$$

Suppose now we look at the structure of Eq. (6). It is of the form

$$(6a) \qquad \Omega_E \circ \psi = E_T \psi$$

Note that E_T is exactly what you think it is: the total energy. We already know that it is a time independent quantity. We also know that it is an x-independent quantity. It is therefore a constant in Eq. (6a). $\Omega_E \circ$ is the energy operator. The structure is therefore

operator working on ψ = constant times ψ

We call this type of equation an eigenvalue equation. In particular:

E_T is the eigenvalue of Ω_E

and ψ is the eigenfunction of Ω_E

both quantities, E_T and ψ are unknowns. They are determined from Eq. (6) and the boundary values of the problem. But there are other things to which Schroedinger's equation must supply the answer. We have already seen the emergence and importance of the dispersion relationship. We derived it for a free particle. But Schroedinger's equation allows the potential energy to be a function of x. The restriction of a free particle is removed and, therefore, the dispersion relationship $E = E(k)$ must change. Schroedinger's equation must give the answer to the question: What is $E(k)$ for a given $V(x)$. The existence of the dispersion relationship in turn resulted in the uncertainty principle. It too should be a part of this equation. Finally, the transformation of F = ma to

$$\vec{F} = \hbar \frac{d\vec{k}}{dt} = \frac{m}{\hbar} \bar{\nabla}_k \bar{\nabla}_k E \frac{d\bar{k}}{dt}$$

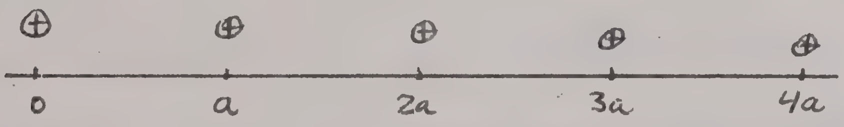
or

$$m = \frac{\hbar^2}{\bar{\nabla}_k \bar{\nabla}_k E}$$

Has to be preserved.

SCHROEDINGER'S EQUATION IN THE CRYSTAL

In order to study the motion of an electron in a crystal let us consider a one-dimensional array of nuclei. Each lattice site carries the charge +q.



The electrostatic potential associated with this array is

$$v(x) = \frac{q}{4\pi\varepsilon x} + \frac{q}{4\pi\varepsilon(x-a)} + \frac{q}{4\pi\varepsilon(x-2a)} + \ldots.$$

If the array is very large so that end effects are small the electrostatic potential will be periodic:

$$v(x) = v(x-a) = v(x-2a) = \ldots..$$

The force on the electron due to this potential is

$$F = (-q)\varepsilon$$

$$= q \frac{dv}{dx}$$

The work associated with moving this charge is

$$W = q(v(x) - v(b))$$

If we choose the position b so that v(b) = 0 the work and therefore the potential energy will be given by

$$V(x) = qv(x)$$

Schroedinger's equation becomes

$$\left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + qv(x)\right)\psi = E\psi$$

Here $v(x)$ is of course periodic. The equation is difficult to solve. We will attempt approximate solutions.

CONSTANT POTENTIAL

Suppose we approximate the potential $V(x)$ by an average value. Then Schroedinger's equation becomes

$$(1) \qquad \left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V_o\right)\psi = E\psi$$

If we assume that ψ has the form

$$(1a) \qquad \psi = A\varepsilon^{jkx} + B\varepsilon^{-jkx}$$

then

$$\frac{d^2\psi}{dx^2} = -k^2\psi$$

Hence, substituting into (1)

$$(2) \qquad E(k) = \frac{\hbar^2 k^2}{2m} + V_o$$

This is the familiar dispersion relationship for a free particle.

The boundary conditions which we would like to use are

$$\left.\begin{array}{ll} (3a) & \psi(x) = \psi(x+na) \\ (3b) & \left.\frac{d\psi}{dx}\right|_x = \left.\frac{d\psi}{dx}\right|_{x+na} \end{array}\right\} n = 0, \pm 1, \pm 2$$

These are periodic boundary conditions. They essentially state that each unit space of the crystal is equally preferred by the electron. Evaluation yields:

$$A\varepsilon^{jkx} + B\varepsilon^{-jkx} = A\varepsilon^{jk(x+na)} + B\varepsilon^{-jk(x+na)}$$

$$jk\ (A\varepsilon^{jkx} - B\varepsilon^{-jkx}) = jk(A\varepsilon^{jk(x+na)} - B\varepsilon^{-jk(x+na)}$$

Addition and subtraction of these two equations yields

$$1 = \varepsilon^{jkna}$$

But $$1 = \varepsilon^{j2\pi m} \qquad m = 0, \pm 1, \pm 2$$

Hence (4) $$k = \frac{2\pi m}{na}$$

Recall that n is also an integer. We fix this integer to a particular value n = N by considering a particular length of crystal L = Na. Then

$$\text{(4a)} \qquad k = \frac{2\pi m}{L} \quad m = 0, \pm 1, \pm 2$$

This suggests that only discrete values of k and therefore E are solutions of the differential equation and the boundary condition.

In order to determine ψ we note that in view of (4a) equation (1a) may be rewritten as

$$\text{(1b)} \qquad \psi = C\varepsilon^{jkx}$$

The coefficient C is set by the normalization condition. We would like to say that the electron is somewhere in the interval $0 \leq x \leq L$. Hence:

$$\int_0^L \psi^*\psi \, dx = C^*C \;\; L = 1$$

$$\sqrt{C^*C} = \frac{1}{\sqrt{L}}$$

or

$$\text{(1c)} \qquad \psi = \frac{1}{\sqrt{L}} \varepsilon^{jkx} \qquad k = \frac{2\pi m}{L}, \quad m = 0, \pm 1, \pm 2 \, ..$$

To each value of m belongs a value of E and k and a wavefunction ψ. We say ψ_m describes a state of the electron with quantum number m. We note that we have a dispersion relationship. The form of E(k) agrees with our previous results. Note, however, that we have learned nothing about the effect of the varying potential. We will study that aspect next.

FINITE DIFFERENCE APPROXIMATION

Consider the situation the crystal. The lattice sites are separated by the lattice constant a. We may take advantage of this as follows: Suppose we are at lattice site r and are interested in lattice site r+1. Then we may use a taylor series as follows:

$$f(ra+a) = f(ra) + \frac{df}{dx}\Big|_{ra} a + \frac{d^2f}{dx^2}\Big|_{ra} \frac{a^2}{2} + \ldots.$$

Furthermore

$$f(ra-a) = f(ra) - \frac{df}{dx}\Big|_{ra} a + \frac{d^2f}{dx^2}\Big|_{ra} \frac{a^2}{2} + \ldots.$$

Addition yields

$$(1) \quad \frac{1}{a^2} [f((r+1)a) + f((r-1)a) - 2f(ra)] = \frac{d^2f}{dx^2}\Big|_{ra}$$

This is the finite difference approximation for the second derivative evaluated at $x = ra$. Consider Schroedinger's equation:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi + V(x)\,\psi = E\psi$$

If we evaluate this at $x = ra$ we obtain

$$(2) \qquad -\frac{\hbar^2}{2ma^2}[\psi(r+1) + \psi(r-1) - 2\psi(r)] + V(r)\,\psi(r) = E\psi(r)$$

Note that this equation is evaluated at $x = ra$ so that $V(ra)$ is a constant. We now shift our energy reference so that $V(ra) = 0$. Then

$$(3) \qquad \psi(r+1) + \psi(r-1) - 2\psi(r) + \frac{2ma^2}{\hbar^2}E\psi(r) = 0$$

or

$$(3a) \qquad \psi(r+1) + \psi(r-1) - [2 - (\frac{a}{r_B})^2 \frac{E}{E_\infty}]\,\psi(r) = 0$$

Here r_B is the lowest Bohr radius (.529A°) and E_∞ is the ionization energy for hydrogen (13.6eV).

We already know the solution for ψ for a constant potential. We will use it here to assume that the solution is of the form

$$(4) \qquad \psi = \sum_k C_k\, \varepsilon^{jkx} \qquad k = \frac{2\pi m}{L}$$

then

$$\psi(ra) = \sum_k C_k \, \varepsilon^{jkra} = \psi(r)$$

$$\psi((r+1)a) = \sum_k C_k \, \varepsilon^{jka(r+1)} = \varepsilon^{jka} \, \psi(r)$$

$$\psi((r-1)a) = \sum_k C_k \, \varepsilon^{jka(r-1)} = \varepsilon^{-jka} \, \psi(r)$$

Substitution into (3a) yields

$$\text{(3b)} \qquad \sum_k C_k \left[\varepsilon^{jka} + \varepsilon^{-jka} - \left[2 - \left(\frac{a}{r_B}\right)^2 \frac{E}{E_\infty} \right] \right] \varepsilon^{jkra} = 0$$

One can now show that the only condition under which (3b) can be satisfied is

$$\text{(5)} \qquad \varepsilon^{jka} + \varepsilon^{-jka} - \left[2 - \left(\frac{a}{r_B}\right)^2 \frac{E}{E_\infty} \right] = 0$$

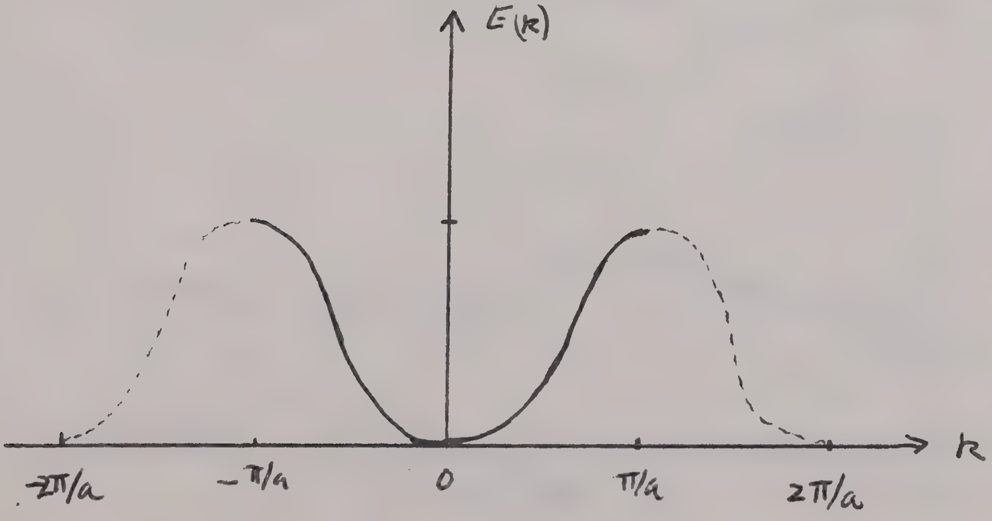
But then

$$E = 2E_\infty \left(\frac{r_B}{a}\right)^2 \cos(ka-1)$$

or

$$\text{(6)} \qquad E(k) = E_\infty \left(\frac{2r_B}{a}\right)^2 \sin^2 \frac{ka}{2}$$

Notice that this equation is our dispersion relationship. It is quite different from our previous result.



An energy band is formed. The width of this band is

$$(7) \qquad E_w = E_\infty \left(\frac{2r_B}{a}\right)^2$$

if the crystal constant is around 2Å this would yield a bandwidth E_w near 4 eV. This is a reasonable result for a crystal.

CONDUCTION AND VALENCE BAND

In the previous section we considered the finite difference expansion of Schroedinger's equation in terms of neighboring lattice points. Let us modify this expansion as follows:

(1) We expand about x = ra to the points $(r \pm \frac{1}{2})a$

(2) We expand about $x = (r + \frac{1}{2})a$ to the points ra and (r+1) a.

The resulting equations are

$$(1) \qquad -\frac{2\hbar^2}{ma^2} \left[\psi(r + \tfrac{1}{2}) + \psi(r - \tfrac{1}{2}) - 2\psi(r)\right] + V(r)\psi(r) = E\,\psi(r)$$

$$(2) \qquad -\frac{2\hbar^2}{ma^2} \left[\psi(r) + \psi(rH) - 2\psi(r+ \tfrac{1}{2})\right] + V(r+ \tfrac{1}{2})\psi(r+ \tfrac{1}{2}) = E\,\psi(r+ \tfrac{1}{2})$$

we now assume that

$$(3) \qquad \psi(r) = \sum_k C_{ok}\, \varepsilon^{jkar} \qquad \psi(r+1) = \varepsilon^{jka}\, \psi(r)$$

$$(4) \qquad \psi(r+ \tfrac{1}{2}) = \sum_k C_{\frac{1}{2}k}\, \varepsilon^{jka(r+ \frac{1}{2})} \qquad \psi(r- \tfrac{1}{2}) = \varepsilon^{-jka}\, \psi(r+ \tfrac{1}{2})$$

We define $E_a = \left(\frac{2r_B}{a}\right)^2 E_\infty$

Then

$$(1a) \qquad \sum_k \left[C_{\frac{1}{2}k}(1+\varepsilon^{-jka})\varepsilon^{jka(r+ \frac{1}{2})} + C_{ok}\left(\frac{E}{E_a} - \frac{V(r)}{E_a} - 2\right)\varepsilon^{jkar}\right] = 0$$

$$(2a) \qquad \sum_k \left[C_{ok}(1+\varepsilon^{jka})\varepsilon^{jkar} + C_{\frac{1}{2}k}\left(\frac{E}{E_a} - \frac{V(r+ \frac{1}{2})}{E_a} - 2\right)\varepsilon^{jka(r+ \frac{1}{2})}\right] = 0$$

We may simplify the algebra slightly by selecting the zero of energy so that $V(r+\frac{1}{2}) = 0$. The condition for a solution is satisfied if the determinant of Eq. (5) vanishes.

$$(5) \qquad 0 = \begin{vmatrix} 1 + \varepsilon^{-jka} & E'-V'-2 \\ E'-2 & 1 + \varepsilon^{jka} \end{vmatrix}$$

$$0 = 2 + \varepsilon^{jka} + \varepsilon^{-jka} + (E'-2)(2-E' + V')$$

We may write this as

$$(E' - 2 - \frac{V'}{2} + \frac{V'}{2})(E'-2 - \frac{V'}{2} - \frac{V'}{2}) = 2(1 + \cos ka)$$

$$(E'-2 - \frac{V'}{2})^2 = (\frac{V'}{2})^2 + 4 \cos^2 \frac{ka}{2}$$

Hence

$$E' = \frac{V'}{2} + 2 \pm \left[(\frac{V'}{2})^2 + 4 \cos^2 \frac{ka}{2} \right]^{\frac{1}{2}}$$

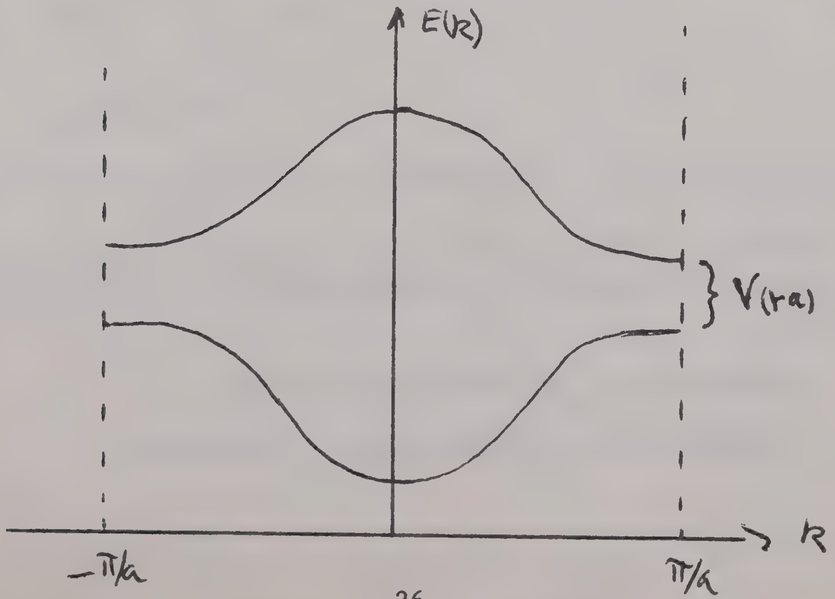
or

$$(6) \qquad E(k) = 2E_a + \frac{1}{2}V(ra) \pm \left[4E_a^2 \cos^2 \frac{ka}{2} + \left| \frac{V(ra)}{2} \right|^2 \right]^{\frac{1}{2}}$$

This is the new dispersion relationship. Note the ± sign on the radical. Both solutions are valid. This would suggest that there are two bands.

$$E(0) = 2E_a + \frac{1}{2} V(ra) \pm [4E_a^2 + (\frac{V(ra)}{2})^2]^{\frac{1}{2}}$$

$$E(\pm \frac{\pi}{a}) = 2E_a + \frac{1}{2} V(ra) \pm \frac{1}{2} V(ra)$$



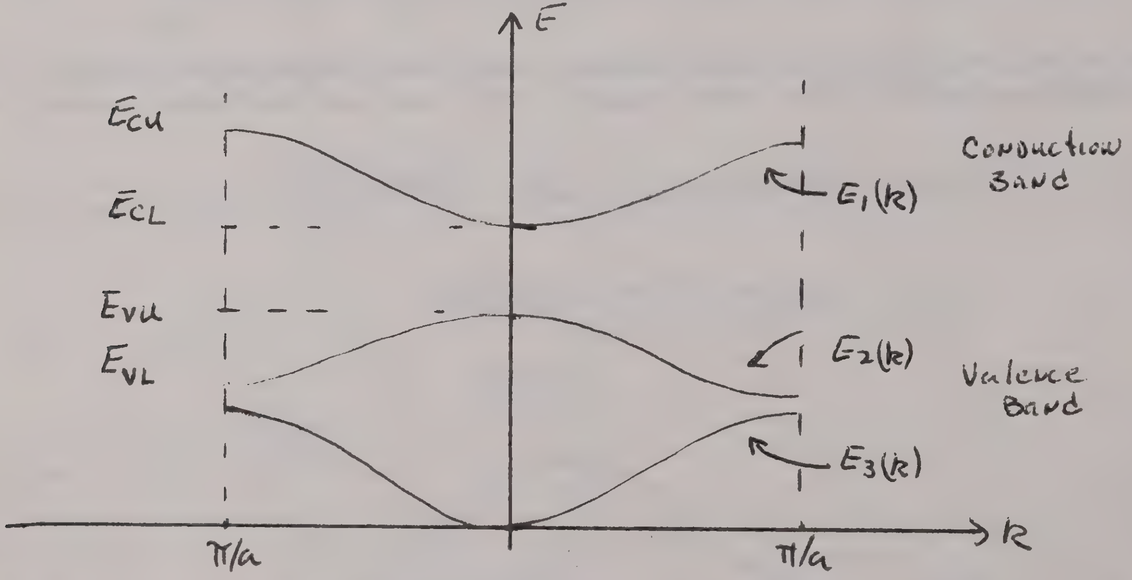
The structure which we are displaying here involves an upper band; the conduction band; and a lower band; the valence band. The two are separated by a region of energies in which we have no states which are described by real values of k. This is the forbidden band. Notice that we have a dispersion relationship in each band.

It would be very interesting to continue our expansion or to attempt a complete solution of Schroedinger's equation. This indeed can be done. The result yields a many band model in which each band has a dispersion relationship E(k). These functions are similar to the functions which are plotted on the previous page. There may, however, be a slight difference in the band structure. This involves the function E(k) at k = 0. E(k) at k = 0 may be a maximum or a minimum. We then have the following situation:

$$E(0) \text{ is maximum, } \left.\frac{dE}{dk}\right|_{k=0} = 0, \quad \left.\frac{d^2E}{dk^2}\right|_{k=0} < 0$$

$$E(0) \text{ is minimum, } \left.\frac{dE}{dk}\right|_{k=0} = 0, \quad \left.\frac{d^2E}{dk^2}\right|_{k=0} > 0$$

In the conduction band for silicon we know that the second situation occurs. This means that $\left.\frac{d^2E}{dk^2}\right|_{k=0} > 0$, so that the effective mass near the bottom of the conduction band is positive. The situation is shown below:



If we want to talk about this model we have to find the wavefunction ψ. This we are not going to do. Instead we assume that the following approximate model is good enough:

(1) Assume that $E_c(k)$ and $E_v(k)$ are known.

(2) Calculate the effective mass in the bands:

$$\frac{1}{m_c} = \frac{1}{\hbar^2} \frac{d^2E_c}{dk^2} \qquad \frac{1}{m_v} = \frac{1}{\hbar^2} \frac{d^2E_v}{dk^2}$$

(3) Approximate m_c and m_v by suitable average values. If, for instance, the conduction band population is small we might say

$$< \frac{1}{m_c} > = \frac{1}{\hbar^2} \left. \frac{d^2E_c}{dk^2} \right|_{k=0}$$

(4) Identify appropriate potential energy in each band which we may use in the constant potential expression:

$$E = \frac{\hbar^2 k^2}{2m^*} + V_o \qquad m^*\text{-effective mass average}$$

in the conduction band $E(k=0) = E_{cL}$, here $V_o = E_{cL}$.

(5) Extend our one dimensional model to three dimensions. Thus

$$\vec{k} = \hat{x}\, k_x + \hat{y}\, k_y + \hat{z}\, k_z$$

and $k_x = \frac{2\pi n_x}{L_x}$, $k_y = \frac{2\pi n_y}{L_y}$, $k_y = \frac{2\pi n_z}{L_z}$

Where the crystal dimensions are $L_x \times L_y \times L_z$ and the three quantum numbers n_x, n_y and n_z occur.

DENSITY OF STATES FUNCTION

Let us now examine the conduction band in some detail. The question which we must answer eventually in order to study conduction phenomena involves the total number of electrons in this band. But an electron is an occupied state. Then the maximum number of electrons which we can have in a given energy range, say $E_{cL} \leq E \leq E_u$, would be precisely the number of states which are available in this energy interval. We know

$$(1) \qquad E_c = \frac{\hbar^2 k^2}{2m_c} + E_{cL}$$

$$(2) \qquad k_x = \frac{2\pi m_x}{L_x} \; , \; k_y = k_y = \frac{2\pi m_y}{L_y} \; , \; k_z = \frac{2\pi m_z}{L_z}$$

This would imply that if we consider k-space one state would occupy the volume

$$(3) \qquad dV_k = \frac{(2\pi)^3}{L_x L_y L_z}$$

But the quantity $L_x L_y L_z = V_{ACT}$; the physical volume of the crystal; is a known quantity. Hence

$$(4) \qquad dV_k = \frac{(2\pi)^3}{V_{ACT}}$$

Consider now equation (1). In k-space it becomes

$$(1a) \qquad E_c - E_{cL} = \frac{\hbar^2}{2m_c} \; (k_x^{\,2} + k_y^{\,2} + k_z^{\,2})$$

Thus, in k-space constant energy surfaces are spheres. The differential spherical shell in this space is

$$(5) \qquad dV_{SPH} = 4\pi \, k^2 \, dk$$

$$k = |\hat{X}k_x + \hat{Y}k_y + \hat{Z}k_z|$$

The number of states in this element are

$$dN_k = \frac{dV_{SPH}}{dV_k} = V_{ACT} \; \frac{1}{2\pi^2} \; k^2 dk$$

The number of states per unit volume of crystal in the k-shell are

$$(6) \qquad dn_k = \frac{k^2}{2\pi^2} \, dk$$

But

$$k = \left(\frac{2m_c}{\hbar^2} (E-E_{cL})\right)^{\frac{1}{2}}$$

Hence

$$dk = \frac{m}{\hbar^2} \left(\frac{2m}{\hbar^2} (E-E_{cL})\right)^{-\frac{1}{2}} dE$$

Substituting into Eq. (6) yields

$$(7) \qquad dn_E = 2 \, \frac{1}{4\pi^2} \left(\frac{2m_c}{\hbar^2}\right)^{3/2} (E-E_{cL})^{+\frac{1}{2}} \, dE$$

The factor of 2 is introduced to account for spin degeneracy. Equation (6) and (7) are of the form

$$(6a) \qquad dn_k = \rho(k) \, dk$$

$$(7a) \qquad dn_E = \rho(E) \, dE$$

We call the two functions $\rho(k)$ and $\rho(E)$ density of state functions in k and E-space respectively. Their meaning is the following: The number of states in the energy interval $[E + dE, E]$ per unit volume of crystal is $\rho(E)dE$. Thus, if we want to compute the number of states per unit volume in $E_{cL} \leq E \leq E_u$ we have

$$N = \frac{2}{4\pi^2} \left(\frac{2m_c}{\hbar^2}\right)^{3/2} \int_{E_{cL}}^{E_{2e}} (E-E_{cL})^{+\frac{1}{2}} \, dE$$

$$= \frac{1}{\pi^2} \left(\frac{2m_c}{\hbar^2}\right)^{3/2} (E_{2e}-E_{cL})^{3/2}$$

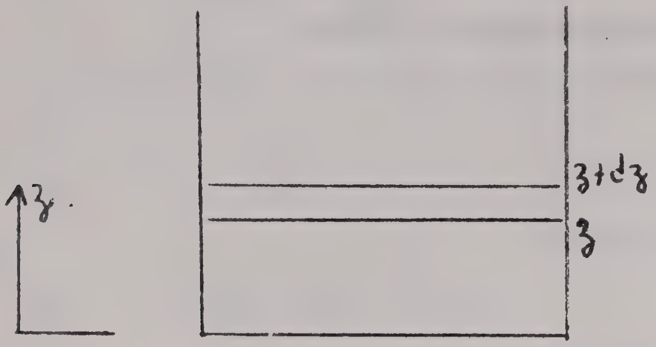
Note that we are talking about states. Nothing has been said about the fact that these states are occupied. If we want to calculate the number of electrons in the conduction band we need the occupation probability for the same at energy E: f(E). Then

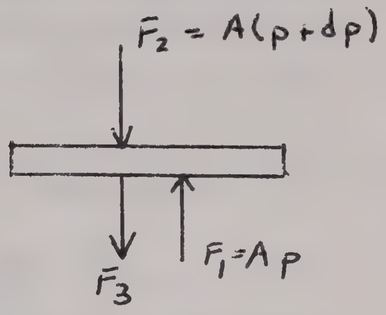
$$(8) \qquad dn_{e-} = \rho(E) \, f(E) \, dE$$

would be the number of electrons in [E+dE,E].

OCCUPATION PROBABILITY

In order to derive an expression for f(E) let us consider a perfect gas as a digression on our main topic:





Consider a section of the ideal gas. Its weight is

$$dW = g\rho A dz = F_3$$

For equilibrium we require

$$\sum_i F_i = 0$$

or

$$g\rho A dz + A(p+dp) - Ap = 0$$

Hence

$$(1) \qquad dp = -g\rho dz$$

But this is a perfect gas:

$$(2) \qquad pV = \alpha RT$$

Here α is the volume mole fraction: $\alpha = \frac{V}{V_m}$, R is the gas constant and T the temperature of the gas. Then

$$(2a) \qquad p\,V_M = RT$$

Let us now ask the question: What is the distribution of molecules/unit volume along the z-axis? In order to do this let us write the mass density ρ as

(3) $$\rho = \frac{mN}{V}$$

m: Mass of a single molecule

N: Number of Molecules in the volume V.

If we take one molar volume N becomes Avagadro's number.

(3a) $$\rho = \frac{mN_0}{V_M}$$

But then by substituting (3a) into (2a):

(2a) $$p = \frac{RT}{V_M} = \frac{RT}{N_0} \frac{\rho}{m}$$

If we define the molecule density by $n = \frac{N}{V}$:

(4) $$p = \frac{RT}{N_0} n$$

Note that $\frac{R}{N_0}$ is a constant which we call k, the Boltzman constant.

Then

(4a) $$p = kT\, n$$

Let us now assume that the entire gas is at temperature T. Then

(4b) $$dp = kT\, dn$$

But we already have an equation for dp: (1). Hence

(1) $$kT\, dn = -g\, mn\, dz$$

or

(5) $$\frac{dn}{n} = -\frac{gm}{kT}\, dz$$

integration yields

(6) $$n(z) = B\varepsilon^{-\frac{mgz}{kT}}$$

This then is the molecular density along the z-axis. Let us investigate the number of molecules along the z-axis:

$$n(z)\ Adz = B\varepsilon^{-\frac{mgz}{kT}}\ Adz$$

$$dN(z) = \frac{AB}{mg}\varepsilon^{-\frac{mgz}{kT}}\ dmgz$$

But mgz is the potential energy of the molecule in the gravitational field.

$$(7) \qquad E = mgz$$

Then

$$(8) \qquad dN(E) = K\ \varepsilon^{-\frac{E}{kT}}\ dE$$

Here K is some constant. Note that $K\ \varepsilon^{-\frac{E}{kT}}$ is the number of molecules in [E+dE,E]. In order to evaluate K let us assume that the total number of molecules in the cylinder is known to be N_T. Then

$$K = \frac{N_T}{kT}$$

Hence

$$(8a) \qquad dN(E) = \frac{N_T}{kT}\ \varepsilon^{-\frac{E}{kT}}\ dE$$

Notice that we can normalize this equation to the volume of the crystal without difficulty.

We are talking about an ideal gas: non-interacting, identical particles. If we try to analyse this by our earlier ideas about states we note the following:

(a) Non-interaction means that a molecule cannot tell if a state is already occupied by another molecule.

(b) This implies that all molecules of the system, N_T, may be in the same state.

(c) The density of state function is of the form

$$\rho(E) = \frac{\text{number of states in the volume } V_E \text{ in E-space}}{\text{size of one state in E-space}}$$

(d) The size of the state was determined from quantization conditions.

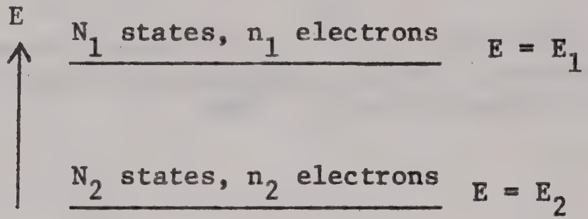
With these ideas in mind we note that (8a) might suggest that the density of state function is of the form $\frac{N_T}{kT}$, which in turn would result in the conclusion that the size of a single state is kT. This is wrong since point (d) does not exist for this classical system. Nevertheless, we might think of an equivalent, non-classical system which has this property. If we do this we identify the function

$$(9) \qquad f(E) = \varepsilon^{-\frac{E}{kT}}$$

as the occupation probability for the states which are located at E. This probability applies to non-interacting systems.

THE TWO LEVEL PROBLEM

Let us consider a system of states for non-interacting particles or electrons which looks as follows:



This system is in equilibrium in the sense that the electron population is constant in time. Thus:

$$n_1 = f(E_1)\, N_1 = \varepsilon^{-\frac{E_1}{kT}} N_1$$

$$n_2 = f(E_2)\, N_2 = \varepsilon^{-\frac{E_2}{kT}} N_2$$

Let us now assume that the two energy levels communicate with each other in the sense that electrons transfer from #1 to #2 and #2 to #1. The transitions are assumed to be proportional to the electron population in i and the number of states in j. Hence

$$R_{2\to 1} = r_{21}\, n_2\, N_1$$

$$R_{1\to 2} = r_{12}\, n_1\, N_2$$

Here r_{21} and r_{12} are system constants in the sense that A resistor is a system constant in a network. Notice that if $n_1 \neq n_1(t)$ and $n_2 \neq n_2(t)$ then

$$R_{2\to 1} = R_{1\to 2}$$

or
$$r_{21}\, n_2\, N_1 = r_{12}\, n_1\, N_2$$

or $$r_{21}\, \varepsilon^{-\frac{E_2}{kT}} = r_{12}\, \varepsilon^{-\frac{E_1}{kT}}$$

or $$r_{12} = r_{12}\, \varepsilon^{+\frac{E_2-E_1}{kT}}$$

Thus only one of the system constants is unknown. Let us now stay with the same system but change the transition rules. This would mean that r_{12} and r_{21} would remain the same. The new rule is: A transition from i→j if i is occupied and j is empty and is proportional to the full states in i and the empty states in j. Then

$$R_{2\to1} = r_{21}\, n_2\, (N_1 - n_1)$$

$$R_{1\to2} = r_{21}\, n_1\, (N_2 - n_2)$$

Note that we are talking about interacting particles. The form of f(E) must change. Let us try to determine it. At equilibrium

$$r_{21}\, n_2\, (N_1 - n_1) = r_{12}\, n_1\, (N_2 - n_2)$$

$$r_{21}\, f(E_2)(1-f(E_1)) = r_{12}\, f(E_1)(1-f(E_2))$$

$$\frac{r_{21}}{r_{12}}\, \frac{f(E_2)}{1-f(E_2)} = \frac{f(E_1)}{1-f(E_1)}$$

But we know the ratio $\frac{r_{21}}{r_{12}}$ to be $\varepsilon^{+\frac{E_2-E_1}{kT}}$. Then

$$\frac{f(E_2)}{1-f(E_2)}\, \varepsilon^{+\frac{E_2}{kT}} = \frac{f(E_1)}{1-f(E_1)}\, \varepsilon^{+\frac{E_1}{kT}}$$

Let us now increase the number of levels. Then we get an equation as shown above. This type of equation can only be satisfied if

$$\frac{f(E_2)}{1-f(E_2)}\, \varepsilon^{+\frac{E_2}{kT}} = K = \frac{f(E_i)}{1-f(E_i)}\, \varepsilon^{+\frac{E_i}{kT}}$$

or $$f(E_i) = \frac{1}{1+ \frac{1}{K} \varepsilon^{\frac{E_i}{kT}}}$$

Let us assume for a minute that we can write K as

$$K = \varepsilon^{\frac{E_F}{kT}}$$

Then, since E_i could be any value of E, we write

$$f(E) = \frac{1}{1+\varepsilon^{\frac{E-E_F}{kT}}}$$

This is the form of the occupation probability function which applies to particles which interact in the filled-empty state manner or which interact via the pauli exclusion principle. It would appear reasonable to state that this interaction would be very weak if the number of states exceeds the number of electrons by a large margin. But this implies that our f(E) should approach the earlier form if $f(t) << 1$. This happens if $E >> E_F$, $T \neq 0$. In that case

$$\begin{aligned} f(E) &\approx \varepsilon^{-\frac{E-E_F}{kT}} \\ &= (\varepsilon^{\frac{E_F}{kT}})\ \varepsilon^{-\frac{E}{kT}} \\ &= K\varepsilon^{-\frac{E}{kT}} \end{aligned}$$

But this is exactly our answer for non-interacting particles. But this would suggest that the quantity $\frac{E_F}{kT}$ is nothing more than a normalization constant which depends on the temperature of the system. We define

(a) E_F to be the fermi energy

and refer to

(b) $f(E) = \frac{1}{1+\varepsilon^{\frac{E-E_F}{kT}}}$ as the Fermi-Dirac-Sommerfeld distribution function

(c) $f_B(E) = A\varepsilon^{-\frac{E}{kT}}$ as the Boltzmann distribution function.

$$\left.\begin{aligned} &\text{We note that } f(E) \rightarrow f_B(E) \\ &A = \varepsilon^{\frac{E_F}{kT}} \end{aligned}\right\} \text{ if } E >> E_F,\ T \neq 0$$

We also note that

$$f(E_F) = \frac{1}{2}$$

Hence $f > \frac{1}{2}$ if $E < E_F$

$$f < \frac{1}{2} \text{ if } E > E_F$$

ELECTRON POPULATION IN THE CONDUCTION BAND

The electron population in the band is given by

$$n = \int_{E_{cL}}^{E_{cu}} \rho_c(E)\, f(E)\, dE$$

where E_{cL} is the lower edge of the band and E_{cu} is the upper edge of the band.

In order to evaluate this quantity consider

$$(1) \qquad \rho_c(E) = \frac{1}{2\pi^2} \left(\frac{2m_c}{\hbar^2}\right)^{3/2} (E-E_{cL})^{\frac{1}{2}}$$

Here m_c is the effective mass which we calculate from the correct relationship E(k) in the band:

$$(2) \qquad m_c = \frac{\hbar^2}{\frac{d^2E}{dk^2}}$$

But this is normally a function of E. But in (1) we assumed that m_c is a constant. It would appear then that what we mean by m_c is m_c as written in (2) evaluated as some value of E. We take this as E_{cL}. Hence

$$(2a) \qquad m_c = \left.\frac{\hbar^2}{\frac{d^2E}{dk^2}}\right|_{E=E_{cL}}$$

then

$$(3) \qquad n = \frac{1}{2\pi^2} \left(\frac{2m_c}{\hbar^2}\right)^{3/2} \int_{E_{cL}}^{E_{cu}} \frac{(E-E_{cL})^{1/2}}{1+\varepsilon^{\frac{E-E_F}{kT}}}\, dE$$

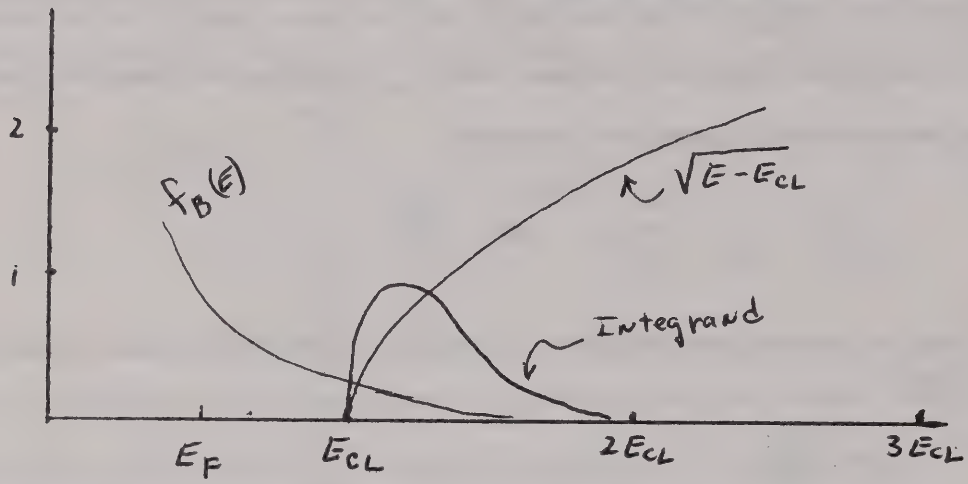
Recall that the conduction band is the outer band. It is nearly empty. But this means that $f(E) << 1$ for some values of $E_{cL} < E < E_{cu}$. But in order for this to be true $E > E_F$. Hence if the band is lightly occupied everywhere

$$(4) \qquad E_{cL} > E_F$$

This suggests that the fermi level is below the conduction band. The condition of sparse population may be used to approximate the integral in (3) by

$$(3a) \qquad n = \frac{1}{2\pi^2} \left(\frac{2m_c}{\hbar^2}\right)^{3/2} \int_{E_{cL}}^{E_{cu}} (E-E_{cL})^{\frac{1}{2}} \varepsilon^{-\frac{E-E_F}{kT}} dE$$

The integrand may be visualized by sketching the behavior of the two functions.



Note that the integrand is essentially zero near E_{cu}. This would suggest that for purposes of integration one night replace E_{cu} by infinity. Re-writing (3a) yields:

$$(3b) \qquad n = \frac{1}{2\pi^2} \left(\frac{2m_c kT}{\hbar^2}\right)^{3/2} \varepsilon^{-\frac{E_{cL}-E_F}{kT}} \int_0^\infty \left(\frac{E-E_{cL}}{kT}\right)^{\frac{1}{2}} \varepsilon^{-\frac{E-E_{cL}}{kT}} d\,\frac{E-E_{cL}}{kT}$$

The integral in this expression is a number: $\frac{\sqrt{\pi}}{2}$. Hence

$$(5) \qquad n = N_c \varepsilon^{-\frac{E_{cL}-E_F}{kT}} \qquad N_c = \frac{1}{2\pi^2} \left(\frac{2m_c kT}{\hbar^2}\right)^{3/2} \frac{\sqrt{\pi}}{2}$$

Notice that m_c is a material constant. Hence N_c is a material parameter. E_F is not known. This is reasonable in view of our interpretation of this constant.

$$\left(\frac{-\hbar^2}{2m}\frac{d^2}{dx^2}+V_0\right)\psi = E\psi \qquad \text{(Electron)}$$

If m is negative we may think of a particle with negative mass in the sense that it is a missing particle with associated negative energy.

$$\left(-\frac{\hbar^2}{2(-m_v)}\frac{d^2}{dx^2}+V_0\right)\psi = E\psi \qquad \text{(Neg. mass electron)}$$

or

$$\left(-\frac{\hbar^2}{2m_v}\frac{d^2}{dx^2}+(-V_0)\right)\psi = (-E)\psi$$

and

$$\left(-\frac{\hbar^2}{2m_v}\frac{d^2}{dx^2}+\hat{V}_0\right)\psi = \hat{E}\psi \qquad \hat{V}_0 = -V_0,\ \hat{E} = -E$$

This "absence" particle has positive mass. Its energy, on the electron scale is increasing in the negative E direction. Its potential energy is E_{vu}. Its density of state function becomes in terms of negative electron energy $\hat{E}$

$$\rho_v(\hat{E}) = \frac{1}{2\pi^2}\left(\frac{2m_v}{\hbar^2}\right)^{3/2}(\hat{E}-\hat{E}_{vu})^{1/2} \qquad (1)$$

where

$$m_v = +\frac{\hbar^2}{\left.\dfrac{d^2\hat{E}}{dk^2}\right|_{\hat{E}=\hat{E}}} \qquad (2)$$

The occupation probability for the absence of the electron is

$$F(E) = 1 - f(E) = \frac{1}{1+\varepsilon^{\frac{E_F-E}{kT}}} \qquad (3)$$

But this equation refers to the electron energy, whereas Eq. (1) refers to the negative electron energy. Hence

$$F(\hat{E}) = \frac{1}{1+\varepsilon^{\frac{\hat{E}-\hat{E}_F}{kT}}} \qquad (3a)$$

This occupation probability will be small if $\hat{E} \gg \hat{E}_F$ or $E_F \gg E$. In terms of the valence band this becomes $E_F >> E_{vu}$. The number of absent electrons in the valence band, or holes is given by

$$(4) \qquad p = \frac{1}{2\pi^2} \left(\frac{2m_v}{\hbar^2}\right)^{3/2} \int_{\hat{E}_{vu}}^{\hat{E}_{vL}} \frac{(\hat{E}-\hat{E}_{vu})^{\frac{1}{2}}}{1+\varepsilon^{\frac{\hat{E}-\hat{E}_F}{kT}}} \, d\hat{E}$$

or

$$(4a) \qquad p \approx \frac{1}{2\pi^2} \left(\frac{2m_v kT}{\hbar^2}\right)^{3/2} \varepsilon^{-\frac{\hat{E}_{vu}-\hat{E}_F}{kT}} \int_0^{\frac{\hat{E}_{vL}-\hat{E}_{vu}}{kT}} \left(\frac{\hat{E}-\hat{E}_{vu}}{kT}\right) \varepsilon^{-\frac{\hat{E}-\hat{E}_{vu}}{kT}} \, d\hat{E}$$

The integration limit is again extended to infinity. The hole energy is next replaced by the electron energy to obtain

$$(5) \qquad p = N_v \, \varepsilon^{-\frac{E_F-E_{vu}}{kT}} \qquad N_v = \frac{1}{2\pi^2} \left(\frac{2m_v kT}{h^2}\right)^{3/2} \frac{\sqrt{\pi}}{2}$$

in summary then, A hole is a particle with positive mass and positive charge. It represents the absense of the electron. Its energy increases along the negative electron energy direction.

Summary:

We have now shown that the electron and hole population in the two bands is of the form

$$(1)\qquad n = N_c \varepsilon^{-\frac{E_C-E_F}{kT}} \qquad\qquad \frac{n}{N_c} << 1$$

$$(2)\qquad p = N_v \varepsilon^{-\frac{E_F-E_V}{kT}} \qquad\qquad \frac{p}{N_v} << 1$$

Here E_c is the lower edge of the conduction band and E_v is the upper edge of the valence band. The physical significance of N_c and N_v becomes apparent from the following:

Consider a level at $E = E_c$ which has N_c states. The density of state function for the level is

$$(3)\qquad \rho_c(E) = N_c \delta(E-E_c)$$

If Boltzmann statistics are assumed to apply the electron population in the band becomes

$$(4)\qquad n = \int_0^\infty N_c \delta(E-E_c) \varepsilon^{-\frac{E-E_F}{kT}} dE$$
$$= N_c \varepsilon^{-\frac{E_C-E_F}{kT}}$$

The entire conduction band has therefore been replaced by a level with effective state density N_c. A similar argument applies to the valence band. Here N_v is the effective density of states for holes or electrons. The change of sign in energy may now be taken care of by writing

$$(5)\qquad p = \int_0^\infty N_v \delta(E-E_v) \left(1 - \frac{1}{1+\varepsilon^{\frac{E-E_F}{kT}}}\right) dE$$
$$= N_v \frac{1}{1 + \varepsilon^{\frac{E_F-E_V}{kT}}} \approx N_v \varepsilon^{-\frac{E_F-E_V}{kT}}$$

Note that we still have not found an expression for E_F. This is to be expected since E_F depends on some form of normalization, which we have not employed. We introduce it by saying that the semiconductor is charge neutral. But charge dnesity in the material is given by

$$(6) \quad \rho = q(p-n)$$

hence

$$(6b) \quad p = n$$

or

$$N_c \varepsilon^{-\frac{E_c-E_F}{kT}} = N_v \varepsilon^{-\frac{E_F-E_v}{kT}}$$

$$kT \ln N_c - E_c + E_F = kT \ln N_v - E_F + E_v$$

$$(7) \quad E_F = \frac{E_v + E_c}{2} + \frac{kT}{2} \ln \frac{N_v}{N_c}$$

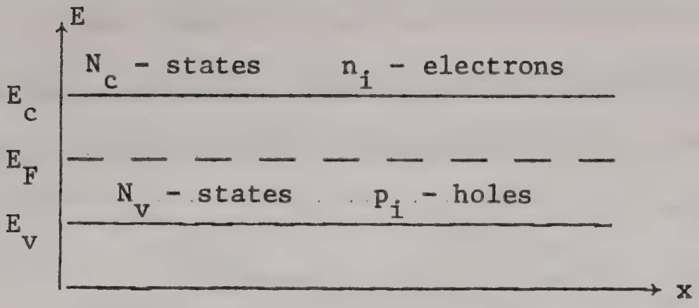
$$(7a) \quad E_F = \frac{E_v + E_c}{2} + \frac{3}{4} kT \ln \frac{m_v}{m_c}$$

This fixes the position of the fermilevel. We may now substitude (7) into (6a) to find the number of electrons or holes

$$n = N_c \varepsilon^{-\frac{E_c - \frac{E_v+E_c}{2} - \frac{kT}{2} \ln \frac{N_v}{N_c}}{kT}}$$

$$(8) \quad n = \sqrt{N_c N_v}\, \varepsilon^{-\frac{E_c-E_v}{2kT}}$$

The quantity $E_c - E_v = \Delta E$ is the width of the forbidden band. It is a material constant (Si - 1.2 eV, Ge - .72 eV). m_v and m_c are material constants. Hence $n = n_i = p$ is a material constant. We call this structure (n = p) an intrinsic semiconductor. Its energy level diagram is shown below.



Extrinsic Semiconductor

Consider the introduction of phosphor into silicon. Since silicon has a valence of 4 and phosphor 5 we have one extra electron. The phosphor atom in the silicon matrix behaves very much like a hydrogen atom: a neutral core and a single outside electron. The atom can be stripped of the electron very easily since the bonding force to the core is weakened by the high dielectric constant of silicon or germanium. ($\varepsilon_R(\text{Si}) = 12$, $\varepsilon_R(\text{Ge}) = 16$.) One may show that the ionization energy is roughly

$$E_i \approx \left(\frac{1}{\varepsilon_r}\right) E_\infty \qquad E_\infty = 13.6 \text{ eV}$$

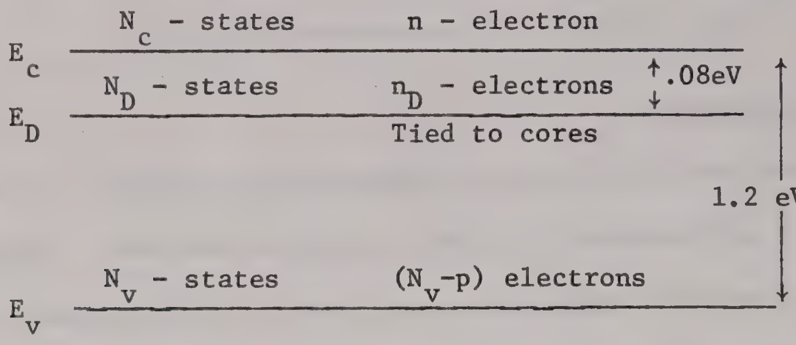
This would suggest that we need roughly 80×10^{-3} eV to ionize the phosphor atom in the silicon lattice.

The energy level in which the electron resides when the phosphor is not ionized must be some 80 meV below the conduction band or above the valence band. This follows since we have no other allowed states in our energy diagram. Suppose we look at the two possibilities. The ionization rates would be

$$R_{D\to v} = r_{Dv}\, n_D p$$

$$\text{or} \qquad R_{D\to c} = r_{DC}\, n_D (N_c - n)$$

If we assume that r_{DV} and r_{DC} are roughly the same, the ionization rate from $D \to c$ will be very much larger and therefore easier than the rate $D \to v$. It would then appear to be more reasonable that the donor level is just below the conduction band, where we use the word "donor" to describe the electron transfer to the conduction band. If the phosphor distribution is uniform we have the following situation.



Note that N_D is the number of phosphor atoms which we have introduced into the silicon. The ionization process is

$$N_D \rightarrow N_D^+ + n_D \uparrow$$

Thus, n_D, the number of electrons at E_D represents the unionized donors. The ionization results in a fixed immobile ion and a conduction band electron, which is free to move.

In order to find the fermi energy we again utilize the charge neutrality concept:

$$(1) \qquad 0 = q(p-n+N_D^+)$$

Since the donors are represented by a level, we know the density of state function at E_D. Ionization represents the absence of an electron at the donor level. Hence

$$(2) \qquad N_D^+ = \int_0^\infty N_D \delta(E-E_D)\big(1-f(E)\big)\, dE$$

$$= \frac{N_D}{1 + \varepsilon^{\frac{E_F-E_D}{kT}}}$$

All donors will be ionized if $\varepsilon^{\frac{E_F-E_D}{kT}} << 1$ or $E_D > E_F$. Conversely, if $E_F > E_D$ none of the donors will be ionized. We would take these statements to mean that the system changes behavior with temperature.

The expression for holes and electrons is

$$n = N_c\varepsilon^{-\frac{E_c-E_F}{kT}} \qquad p = N_v\varepsilon^{-\frac{E_F-E_v}{kT}}$$

The product of these two quantities is

$$(3) \qquad pn = N_cN_v\varepsilon^{-\frac{E_c-E_v}{kT}} = n_i^2$$

We use (3) to re-write (1)

$$0 = \frac{n_i^2}{n} - n + \frac{N_D}{1 + \varepsilon^{-\frac{E_c-E_F}{kT}}\varepsilon^{\frac{E_c-E_D}{kT}}}$$

or

$$(4) \qquad 0 = \frac{n_i^2}{n} - n + \frac{N_D}{1 + \frac{n}{n_c}\varepsilon^{+\frac{E_c-E_D}{kT}}}$$

Re-arrangement yields

$$(4a) \qquad 0 = (n_i^2-n^2)\left(1 + \frac{n}{N_c}\varepsilon^{\frac{E_c-E_D}{kT}}\right) + nN_D$$

This equation is a cubic in n. We will examine it by approximations.

(a) Low temperature.

$$\varepsilon^{\frac{E_c-E_D}{kT}} >> 1$$

$$\text{assume} \quad \frac{n}{N_c}\varepsilon^{\frac{E_c-E_D}{kT}} >> 1$$

$$(4b) \qquad n^2 - n_i^2 \approx N_cN_D\varepsilon^{-\frac{E_c-E_D}{kT}}$$

$$n \qquad N_cN_v\varepsilon^{-\frac{E_c-E_v}{kT}} + N_cN_D\varepsilon^{-\frac{E_c-E_D}{kT}}$$

$$(4c)\quad n \approx \sqrt{N_c N_D}\ \varepsilon^{-\frac{E_c - E_D}{2kT}}$$

(b) Large temperature.

$$\varepsilon^{\frac{E_c - E_D}{kT}} \sim 10$$

$$\frac{n}{N_c} < < 1 \quad \text{(always)}$$

$$(4d)\quad n^2 \approx n_i^{\,2} + nN_D$$

$$\left(n - \frac{N_D}{2}\right)^2 = N_c N_v \varepsilon^{-\frac{E_c - E_v}{kT}} + \left(\frac{N_D}{2}\right)^2$$

$$(4e)\quad n \approx \frac{N_D}{2} + \left[N_c N_v \varepsilon^{-\frac{E_c - E_v}{kT}} + \left(\frac{N_D}{2}\right)^2\right]^{1/2}$$

If the temperature is high enough

$$n_i > > N_D$$

This yields (4f) $n \approx n_i$

On the other hand it would appear that our "large" temperature approximation is valid at room temperature. This would suggest that if

$$n_i < < N_D$$

$$(4g)\quad n \approx N_D$$

The temperature at which (4g) must be replaced by (4f) may be estimated by studying the condition

$$n_i = N_D$$

$$\sqrt{N_c N_v}\ \varepsilon^{-\frac{\Delta E}{2kT_u}} = N_D$$

$$\Delta E = 2kT_u \ln \frac{\sqrt{N_c N_v}}{N_D}$$

or

$$(5) \qquad \frac{T_u}{T_{300}} = \frac{\Delta E}{2kT_{300}} \; \frac{1}{\ln \dfrac{\sqrt{N_c N_v}}{N_D}}$$

Above T_u we use $n \approx n_i$, below T_u we use $n \approx N_D$. Evaluation for silicon yields

$$(5a) \qquad \frac{T_u}{T_{300}} = \frac{22}{\ln \dfrac{10^{19}}{N_D}}$$

Note that T_u increases with N_D, the doping level. This would suggest that the physical significance of T_u is nothing more than the swamping out of the electron contribution due to the donors by the intrinsic carriers, which are increasing rapidly with temperature.

We may find a similar corner temperature for the low temperature border. This is accomplished via (4c) and (4g).

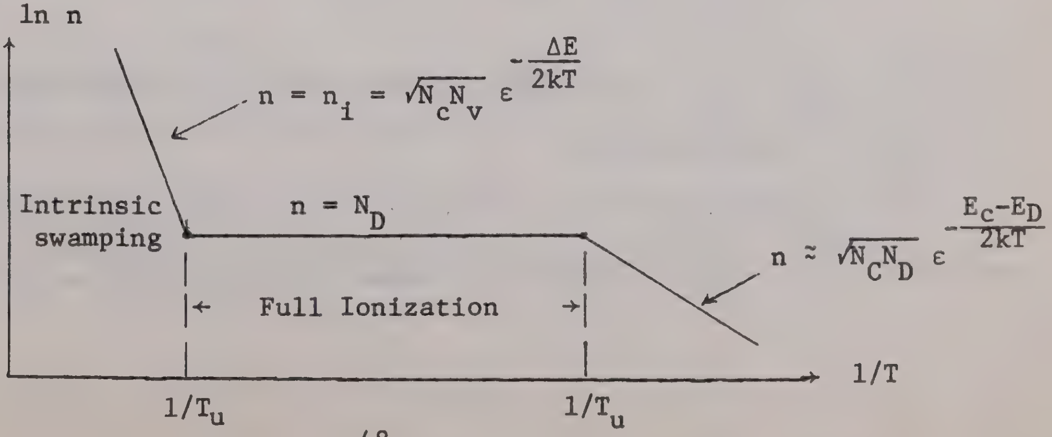
$$N_D = \sqrt{N_c N_D} \; \varepsilon^{-\frac{E_c - E_D}{2kT_L}}$$

or

$$(6) \qquad \frac{T_L}{T_{300}} = \frac{E_c - E_D}{kT_{300}} \; \frac{1}{\ln \dfrac{N_c}{N_D}}$$

The physical significance here involves the inpurity itself. If $T > T_L$ nearly all impurities are ionized, if $T < T_L$ we speak of freeze out.

We summarize our results in the following figure:



The addition of the impurity alters the electrical behavior of the semiconductor. It was added to obtain a definite effect: $n \neq p$. Note that this condition is not met for $T > T_u$. It is met only marginally for $T < T_L$. The impurity is most effective for $T_L < T < T_u$, which defines the operating range of the doped material. In this range

$$n \approx N_D \qquad\qquad n_i << N_D$$

$$p = \frac{n_i^2}{n} << n_i$$

The fermi level is easily calculated from

$$n = N_D = N_c^{\frac{E_c-E_F}{kT}}$$

or

$$kT \ln \frac{N_D}{N_c} = -(E_c-E_F)$$

$$E_F = E_c - kT \ln \frac{N_c}{N_D}$$

It is sometimes convenient to measure the fermi level with respect to the intrinsic fermi level: $E_{Fi} \approx \frac{E_c + E_v}{2}$

$$n = N_c \varepsilon^{-\frac{E_c-E_F}{kT}} = N_c \varepsilon^{-\frac{E_c-E_{Fi}}{kT}} \varepsilon^{\frac{E_F-E_{Fi}}{kT}}$$

$$n = n_i \varepsilon^{\frac{E_F-E_{Fi}}{kT}}$$

Hence

$$kT \ln \frac{N_D}{n_i} + E_{Fi} = E_F$$

The distance $E_F - E_{Fi}$ is therefore proportional to the logarithm of the normalized impurity concentration.

The semiconductor $n = N_D >> n_i$, $n >> p$ is defined as an N-type semiconductor. If a type III impurity is added to the silicon we may show that

$p = N_A >> n_i$, $p >> n$. This is a p-type semiconductor. The dopant is an acceptor in the sense that it accepts an electron from the valence band when it ionizes, it therefore forms a hole in the valence band and the ion N_A^-. Its energy level is E_A which is a few electron volts above E_v. Its behavior is summarized as:

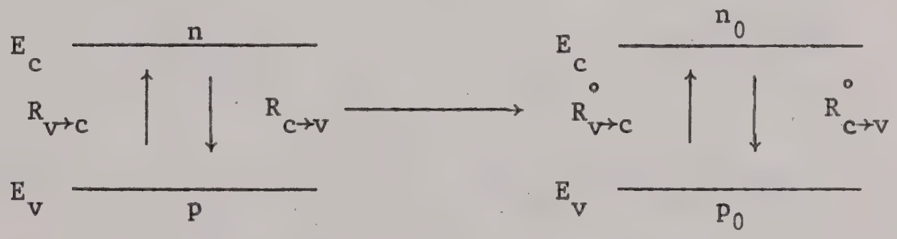
$$p \approx N_A \qquad n = \frac{n_i^2}{N_A} \qquad E_F = E_v + kT \ln \frac{N_A}{N_v}$$

$$= E_{Fi} - kT \ln \frac{N_A}{n_i}$$

Recombination

So far we have considered an equilibrium semiconductor. Suppose we move this restriction as follows: we shine light on the material and turn it off at say t = 0. The question becomes: how does the system return to equilibrium?

In order to analyze this condition let us assume that for $t < 0$ the hole and electron population is p and n. This population returns to p_0 and n_0 after a reasonable period of time. Our system looks as follows:



$$R_{v\to c} = r_{vc}(N_v - p)(N_c - n) \qquad \overset{\circ}{R}_{v\to c} = \overset{\circ}{r}_{vc}(N_v - p_0)(N_c - n_0)$$

$$R_{c\to v} = r_{cv}np \qquad \overset{\circ}{R}_{c\to v} = \overset{\circ}{r}_{cv}n_0p_0$$

We write the non-equilibrium equation at t = 0, i.e., right after the light has been turned off. This would imply $r_{vs} = \overset{\circ}{r}_{vc}$, $r_{cv} = \overset{\circ}{r}_{cv}$. Furthermore, we insist that the population is small compared to the states. Then

$$R_{v\to c} = \mathring{r}_{vc} N_v N_c \qquad \mathring{R}_{v\to c} = \mathring{r}_{vc} N_v N_c$$

$$R_{c\to v} = \mathring{r}_{cv} np \qquad \mathring{R}_{c\to v} = \mathring{r}_{cv} n_0 p_0$$

Note that $R_{v\to c} = \mathring{R}_{v\to c} = \mathring{r}_{vc} N_v N_c$ is the same in both cases. This transition generates an electron hole pair. It is called the thermal generation rate.

At equilibrium:

$$\mathring{r}_{vc} N_v N_c = \mathring{r}_{cv} n_0 p_0 = \mathring{r}_{cv} n_i^2$$

$$\mathring{r}_{vc} = \frac{n_i^2}{N_v N_c} \mathring{r}_{cv}$$

$$\mathring{r}_{vc} = \varepsilon^{-\frac{E_c - E_v}{kT}} \mathring{r}_{cv}$$

In the non-equilibrium case this is not true, i.e., $R_{cv} \neq R_{vc}$. The population is a function of time. Let us then write

$$U_{c\to v} = R_{c\to v} - R_{v\to c}$$

The transition c→v destroys an electron hole pair. This means that the electron population is decreasing if $U_{c\to v} > 0$. Hence

$$\frac{dn}{dt} = -U_{c\to v}$$

or

$$-\frac{dn}{dt} = \mathring{r}_{cv} np - \mathring{r}_{vc} N_c N_v$$

$$= \mathring{r}_{cv} (np - n_0 p_0)$$

Let us write

$$n = n_0 + \Delta n$$

$$p = p_0 + \Delta p$$

Note that Δn and Δp are excess densities over equilibrium. Since pair processes only are involved it is reasonable to say $\Delta n = \Delta p = \delta$. Hence

$$np = (n_0+\delta)(p_0+\delta) = n_0 p_0 + \delta(n_0+p_0) + \delta^2$$

and
$$-\frac{d}{dt}\, n = \frac{d}{dt}\,(n_0+\delta) = \frac{d\delta}{dt}$$

Hence
$$-\frac{d}{dt}\,\delta = \overset{\circ}{r}_{cv}\left(\delta(n_0+p_0) + \delta^2\right)$$

If we assume that $\delta < < n_0 + p_0$ this equation is roughly:

$$-\frac{1}{\delta}\frac{d}{dt}\,\delta = \overset{\circ}{r}_{cv}(n_0+p_0)$$

The equation may be integrated:

$$\ln\frac{\delta(t)}{\delta(0)} = -\overset{\circ}{r}_{cv}(n_0+p_0)t$$

Hence

$$\delta(t) = \delta(0)\varepsilon^{-\overset{\circ}{r}_{cv}(n_0+p_0)t}$$

We now define the lifetime by

$$\tau = \frac{1}{\overset{\circ}{r}_{cv}(n_0+p_0)}$$

Hence

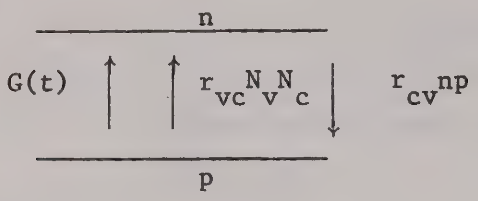
$$n = n_0 + \delta(0)\varepsilon^{-t/\tau}$$

$$p = p_0 + \delta(0)\varepsilon^{-t/\tau}$$

The excess carrier densities decay exponentially. Note again that we have constant thermal generation and a population dependent recombination. Note also that

$$-\frac{dn}{dt} = \quad - G = -\frac{dp}{dt} = \frac{\delta}{\tau}$$

Suppose we look at the generation mechanism somewhat closer. The thermal generation rate is a system constant: $\overset{\circ}{r}_{v\to c}N_cN_v$. How then did the system go to the state, n, p^2. The only possibility appears to be an external generation rate. This external generation rate must occur together with the thermal generation rate.



Note that we can write immediately

$$-\frac{dn}{dt} = -G_{TH} - G_{EXT} = \frac{\delta}{\tau} - G(t) = -\frac{dp}{dt}$$

But then

$$\frac{d\delta}{dt} = -\frac{\delta}{\tau} + G(t)$$

Suppose now that

$$G(t) = G_0U(t)$$

Then the Laplace transform technique yields

$$S\Delta = -\frac{\Delta}{\tau} + \frac{G_0}{S}$$

or

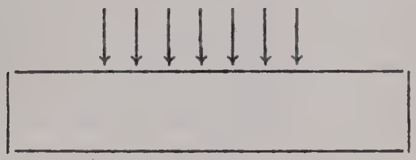
$$\Delta = \frac{G_0}{S\left(S + \frac{1}{\tau}\right)} = G_0\tau\left(\frac{1}{S} - \frac{1}{S + \frac{1}{\tau}}\right)$$

and

$$\delta(t) = G_0\tau(1 - \varepsilon^{-t/\tau})$$

The excess density builds up exponentially. The system goes from n_0, p_0 to the new steady state $n_0 + G_0\tau$, $p_0 + G_0\tau$. Thus, we get a large change $G_0\tau$ if τ is large. But the time constant is τ. Thus a large changes $G_0\tau$ requires a slow response.

Suppose now that we shine the light on a semiconductor bar.



Then the excess density will be a function of position. But electrons and holes are mobile. They will tend to flow away from the high to the low concentration regions. But this implies current flow.

Current Flow

The current flow with which we are familiar is the so-called ohmic or conduction current or current density. Thus, if we have a charge density ρ which moves with velocity ν we have

$$J = \rho\nu \quad \frac{\text{coul}}{\text{cm}^3} \cdot \frac{\text{cm}}{\text{sec}} = \frac{\text{A}}{\text{cm}^2}$$

Moveable densities are $-qn$ and $+qp$ in the semiconductor. Hence

$$J_p = qp\nu$$

$$J_n = (-qn)(-\hat{\nu})$$

The velocity which we are talking about is the component of the total velocity which is in the direction of J_n and J_p. But holes and electrons change direction often due to interactions with crystal imperfections such as charged donors or acceptors. Since their effective mass and charge differs we would expect the "drift" velocities of the two carriers to be different.

$$J_p = qp\nu_p$$

$$J_n = qn\nu_N$$

The two velocities are associated with the field ε which acts onboth charges. We define

$$\mu_n = \frac{\nu_N}{\varepsilon} \qquad \mu_p = \frac{\nu_p}{\varepsilon}$$

Hence

$$J_p = q\mu_p p\varepsilon$$

$$J_n = q\mu_n n\varepsilon$$

and

$$J_{TOTAL} = J_n + J_p$$

The question which now arises concerns the identification of the field E on the energy diagram. Let us write

$$(1) \quad E_F = E_c + kT \ln \frac{n}{N_C}$$

If the system is in equilibrium E_F is a constant. If the system is in equilibrium and the doping is reasonable

$$n = N_D \qquad \text{(N-type)}$$

If we make $N_D(x)$; non-uniform doping; then $n = n(x)$. Hence

$$(2) \quad E_F = E_c(x) + kT \ln \frac{n(x)}{N_c}$$

Differentiation with respect to x yields

$$(2a) \quad 0 = \frac{dE_c}{dx} + kT \frac{1}{n}\frac{dn}{dx}$$

Note that the units of E_c are electron volts. Hence $\frac{1}{q}\frac{dE_c}{dx}$ would be in volts/cm which are the units of $\mathcal{E}$. This makes sense in view of our earlier discussions on crystal coulomb potentials. Let us write

q $$(3) \qquad q\mathcal{E} = \frac{dE_c}{dx}$$

Hence

$$(2a) \qquad 0 = q\mathcal{E} + kT\,\frac{1}{n}\frac{dn}{dx}$$

or $$(2b) \qquad 0 = q\mu_n n\varepsilon + kT\,\mu_n \frac{dn}{dx}$$

Notice that the first term is the drift component of current. The second term is the diffusion current which we write as

$$(4) \qquad J_n^D = (-q)\,D_n \left(-\frac{dn}{dx}\right)$$

where D_n is the diffusion coefficient for electrons in silicon. Combining (2b) and (4) yields

$$(5) \qquad \frac{kT}{q} = \frac{D_n}{\mu_n}$$

and

$$(2c) \qquad 0 = q\mu_n n\varepsilon + qD_n \frac{dn}{dx}$$

By the same argument

$$(2d) \qquad 0 = q\mu_p p\varepsilon - qD_p \frac{dp}{dx}$$

Thus, equilibrium means that the total electron current: diffusion and drift component is zero and that the total hole current is zero. Non-equilibrium means that this balance is upset. Hence

$$(6) \qquad J_n = q\mu_n n\varepsilon + qD_n \frac{dn}{dx}$$

(7) $J_p = q\mu_p p\varepsilon - qD_p \frac{dp}{dx}$

(8) $J_{TOTAL} = J_n + J_p$

Resistor

Suppose now that we have a piece of silicon and dope it uniformly so that

$$n_0 = N_D$$

$$p_0 = \frac{n_i^2}{N_D}$$

Let us now make contact to the material in such a way that the contact maintains the equilibrium densities. This would suggest that even if we go to a non-equilibrium situation we would have

$$n = n_0 \qquad p = p_0$$

Hence

(6a) $J_n = q\mu_n n_0 \varepsilon \qquad \frac{dn}{dx} = \frac{dn_0}{dx} = 0$

(7a) $J_p = q\mu_p p_0 \varepsilon \qquad \frac{dp}{dx} = \frac{dp_0}{dx} = 0$

If the cross-sectional area of the bar is A:

(8a) $I = J_{TOTAL} \cdot A = A(q\mu_n n_0 + q\mu_p p_0)E$

But the field is related to the voltage by

(9) $E = -\frac{dv}{dx}$

Hence

$$I\,dx = -A(q\mu_n n_0 + q\mu_p p_0)dv$$

If the device extends from x = 0 to x = L integration yields

$$(10)\quad IL = -A(q\mu_n n_0 + q\mu_p p_0)\left(V(L) - V(0)\right)$$

But the applied voltage is defined as

$$V_0 = V(0) - V(L)$$

Hence

$$(10a)\quad V_0 + \frac{L}{(q\mu_n n_0 + q\mu_p p_0)A} I$$

It appears that the conductivity of the material is

$$(11)\quad \sigma = q\mu_n n_0 + q\mu_p p_0$$

and that we are talking about a resistor

$$R = \frac{1}{\sigma}\frac{L}{A}$$

The field is defined by

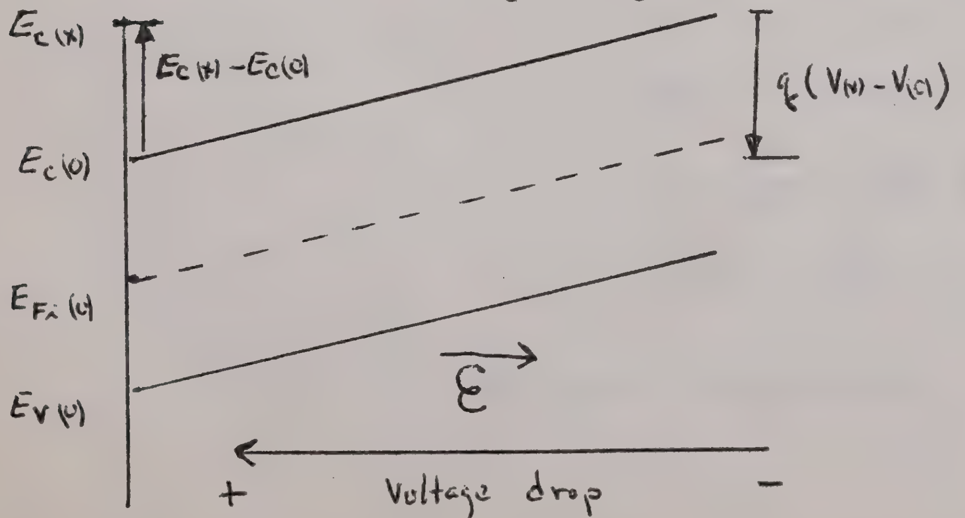
$$\mathcal{E} = \frac{1}{q}\frac{dE_c}{dx} = -\frac{dv}{dx}$$

Hence our representation of voltage on the energy diagram would take the form

$$v(x) - v(0) = -\frac{1}{q}\left(E_c(x) - E_c(0)\right)$$

or

$$-qv(x) + qv(0) = +\left(E_c(x) - E_c(0)\right)$$



Let us summarize our results as follows:

(1) On the energy diagram the slope: $\frac{dE_c}{dx} = \frac{dE_v}{dx} = \frac{dE_{Fi}}{dx}$ represents the field qE.

(2) The distance $E_c(x) - E_c(0)$ represents the negative potential difference $-q[v(x) - v(0)]$. We may state this simply by noting that the most positive point in the energy diagram is the lowest energy point.

(3) In our diagram holes move in the positive x-direction and electrons in the negative x-direction. A useful analogy is: electrons behave like marbles, holes like bubbles on the energy diagram.

(4) In the uniformly doped resistor with non-injecting contacts

$$J = (q\mu_n n_0 + q\mu_p p_0)E$$

$$J = \sigma E$$

$$\frac{dn_0}{dx} = \frac{dp_0}{dx} = 0$$

Suppose now that we illuminate our resistor uniformly. We insist that the contacts remain non-injecting. Then, according to our previous calculations the excess density will be

$$\Delta n = \Delta p = G_0 \tau$$

Hence $$n = n_0 + \Delta n = n_0 + G_0 \tau$$

$$p = p_0 + \Delta p = p_0 + G_0 \tau$$

and $$\frac{dn}{dx} = \frac{dp}{dx} = \tau \frac{dG_0}{dx} = 0$$

If the light G_0 is uniform. Hence

$$J = (q\mu_n n_0 + q\mu_p p_0)E + q(\mu_n + \mu_p)G_0 \tau E$$

or

$$J = \sigma_0 E + \Delta\sigma E$$

The conductivity of the material is being modulated by the light. The effect is called conductivity modulation. Note that if we make a resistor out of the material we get

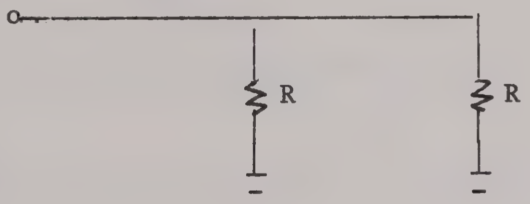
$$R = \frac{1}{\sigma_0 + \Delta\sigma} \frac{L}{A}$$

or

$$\frac{1}{R} = \sigma_0 \frac{A}{L} + \Delta\sigma \frac{A}{L}$$

$$= \frac{1}{R_0} + \frac{1}{R}$$

The equivalent circuit becomes simply:



The requirement on the illumination is only that the illumination is uniform. The light source itself can be time dependent. If this is true the excess density can be calculated from:

$$-\frac{d\delta}{dt} = \frac{\delta}{\tau} - G(t)$$

$$\delta = \Delta n = \Delta p$$

Non-Uniformly Doped Resistor

Nearly all resistors in integrated circuits are produced during the base diffusion process. The diffusion is boron-p-type into a uniform background which is phospor doped. The boron profile is of the form

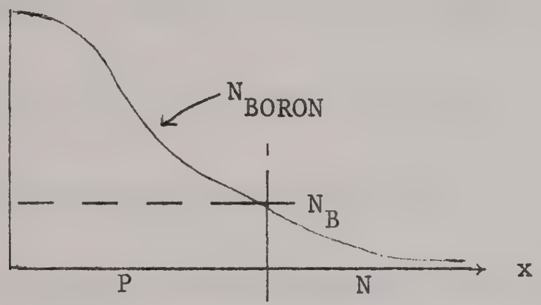
$$N_{BORON} = N_s \varepsilon^{-(ax)^2}$$

N_s - boron surface concentration

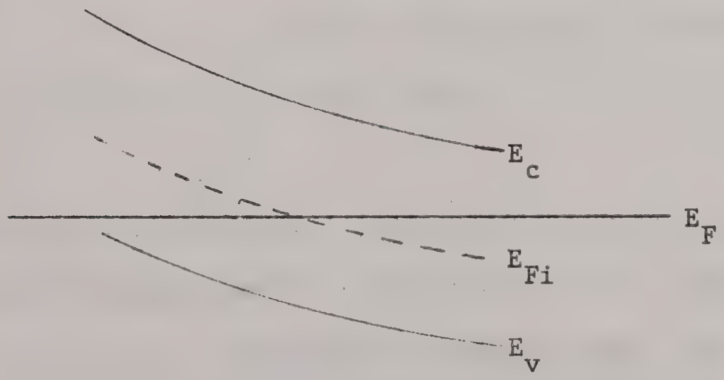
a - constant

The total acceptor impurity is

$$N_A = N_s \varepsilon^{-(ax)^2} - N_B$$



Suppose we assume that the silicon is at equilibrium and that we recall this means E_F = constant. Our energy diagram then looks as follows:



Note that this diagram results in one important conclusion: There is a field without applied voltage, with zero current. We call this the "built-in" field. The question which concerns us here is: How is it evaluated?

The electron and hole currents are individually zero.

$$(1) \quad 0 = q\mu_n n_0 E + qD_n \frac{dn_0}{dx}$$

$$(2) \quad 0 = q\mu_p p_0 E - qD_p \frac{dp_0}{dx}$$

or

$$(1a) \quad E = -\frac{D_n}{\mu_n}\frac{1}{n_0}\frac{dn_0}{dx}$$

$$(2a) \quad E = \frac{D_p}{\mu_p}\frac{1}{p_0}\frac{dp_0}{dx}$$

But $\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{kT}{q}$. Hence

$$(3) \quad E = -\frac{kT}{q}\frac{1}{n_0}\frac{dn_0}{dx} = \frac{kT}{q}\frac{1}{p_0}\frac{dp_0}{dx}$$

But we also know that

$$(4) \quad p_0 n_0 = n_i^2$$

The additional equation which we have used consistently is the charge neutrality equation. The charge density is

$$\rho = q(p_0 - n_0 - N_A^-)$$

This charge density is tied to a field by the equation

$$(5) \quad \overline{\nabla} \cdot \overline{D} = \rho$$

But $\overline{D} = \varepsilon\overline{E}$. Hence, if we have charge neutrality E at best can be a constant. But according to our diagram, E is not a constant. This suggests that we do not have charge neutrality. Our remaining equation becomes:

$$(6) \quad \varepsilon \frac{dE}{dx} = q(p_0 - n_0 - N_A)$$

where we assumed complete ionization. Eq. (3) yields

$$(3a) \quad \frac{dE}{dx} = \frac{kT}{q} \frac{d^2}{dx^2} \ln p_0$$

hence

$$(6a) \quad \frac{\varepsilon kT}{q} \frac{d^2}{dx^2} \ln p_0 = p_0 - \frac{n_i^2}{p_0} - N_A$$

This is the equation from which p_0 is to be determined if $N_A(x)$ is known. The equation is non-linear and has not been solved by analytic means. Its approximate behavior may be studied by

(1) assuming near charge neutrality $(p_0 - \frac{n_i^2}{p_0} - N_A \approx 0)$

or (2) assuming complete non-neutrality $(\rho \approx -q\, N_A)$.

Let us consider condition (1) in some detail. The assumption is

$$p_0 - \frac{n_i^2}{p_0} - N_A = \text{small quantity}$$

If we are assuming reasonable doping: $p_0 >> n_0$ the assumption reduces to

$$p_0 - N_A = \text{small quantity}$$

This would suggest that $p_0 \approx N_A$

We may check this assumption by noting from (6a) that this result would be valid if

$$\frac{\varepsilon kT}{q^2} \frac{d^2}{dx^2} \ln p_0 \bigg|_{p_0 = N_A} = \text{small quantity}$$

but

$$N_A = N_s \varepsilon^{-(ax)^2} - N_B$$

$$\frac{d}{dx} \ln N_A = \frac{-2a^2 x N_s \varepsilon^{-(ax)^2}}{N_s \varepsilon^{-(ax)^2} - N_B}$$

and

$$\frac{d^2}{dx^2}\ln N_A = -\left(\frac{+2a^2xN_s\varepsilon^{-(ax)^2}}{N_s\varepsilon^{-(ax)^2} - N_B}\right)^2 + \frac{(4a^4x^2-2a^2)N_s\varepsilon^{-(ax)^2}}{N_s\varepsilon^{-(ax)^2} - N_B}$$

Inspection of this equation yields the immediate conclusion that our assumption of neutrality is not valid near the junction: $N_A = 0$. Near $x = 0$ we may assume that

$$N_s\varepsilon^{-(ax)^2} >> N_B$$

then

$$\frac{d}{dx}\ln N_A \simeq -2a^2x \qquad \mathcal{E} \simeq -\frac{kT}{q}2a^2x$$

and

$$\frac{d^2}{dx^2}\ln N_A \simeq -2a^2$$

Our requirement for near neutrality becomes now

$$\left|\frac{\varepsilon kT}{q^2}\frac{d^2}{dx^2}\ln N_A\right| << n_i$$

or

$$\frac{\varepsilon kT}{q^2}2a^2 << n_i$$

or

$$2a^2 << \frac{q^2n_i}{\varepsilon kT}$$

Notice that

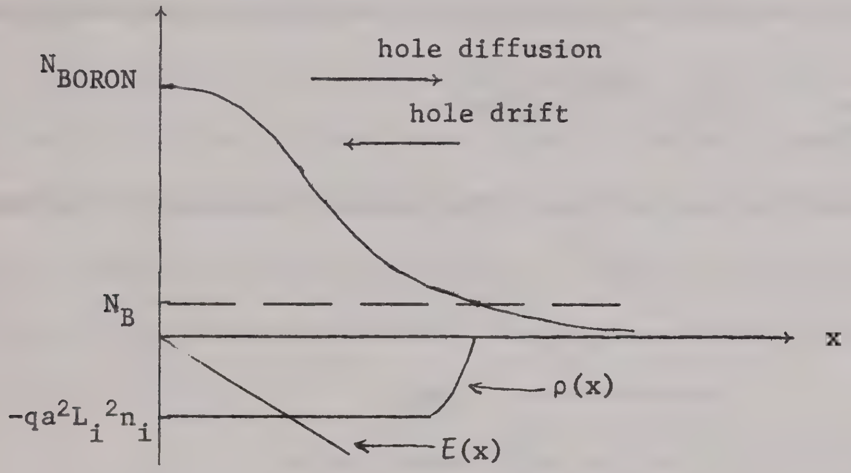
$$\frac{q^2n_i}{\varepsilon kT} = \frac{q\mu_in_i}{\varepsilon D_i} = \frac{\sigma_i}{\varepsilon D_i} = \frac{1}{\tau_iD_i} = \frac{1}{L_i^2}$$

Here L_i is the intrinsic debye length, τ_i the dielectric relaxation time and σ_i the conductivity for intrinsic material. Our criterion ofr near charge neutrality becomes:

$$2a^2L_i^2 << 1$$

This inequality will always be satisfied if the profile changes slowly in one debye length. Indeed, such a criterion may be derived generally, i.e., for an arbitrary profile.

Our development indicates that we have a field which points in the negative x-direction and increases with x. Its origin is simply explained by considering diffusion of majority carriers. This diffusion results in uncompensated acceptors which are the source for the field which limits further diffusion. The uncompensated charge takes the following form.



Summary:

Quasi-neutral case

$$p_0 \approx N_A(x) \qquad \text{or} \qquad n_0 \approx N_D(x)$$

$$n_0 = \frac{n_i^2}{p_0} \qquad\qquad p_0 = \frac{n_i^2}{n_0}$$

$$E = -\frac{kT}{q}\frac{d}{dx}\ln N_A \qquad\qquad E = \frac{kT}{q}\frac{d}{dx}\ln N_D$$

We have already seen that the quasi-neutral approximation fails near the junction. A little thought will reveal that this is very reasonable. Consider the impurity concentration:

$$N_A = N_s \varepsilon^{-(ax)^2} - N_B$$

The junction position is given

$$x_j = \frac{1}{a}\left[\ln \frac{N_s}{N_B}\right]^{1/2}$$

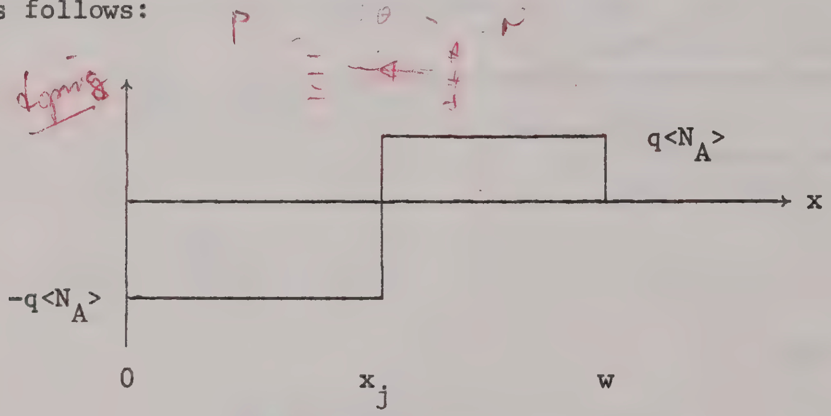
Let us replace the doping profile to the left of the junction by a uniform acceptor density

$$<N_A> = \frac{1}{x_j} \int_0^{x_j} (N_s \varepsilon^{-(ax)^2} - N_B)\,dx$$

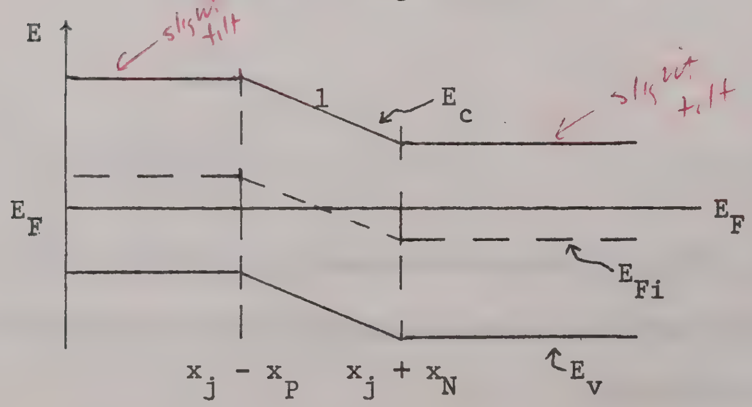
On the right side of the junction we replace the non-uniform doped N-section by

$$<N_D> = \frac{1}{w - x_j} \int_{x_j}^{w} - (N_s \varepsilon^{-(ax)^2} - N_B)\,dx$$

Here w is the total length of the silicon. The junction and energy diagrams are as follows:



Fixed Charge Distribution



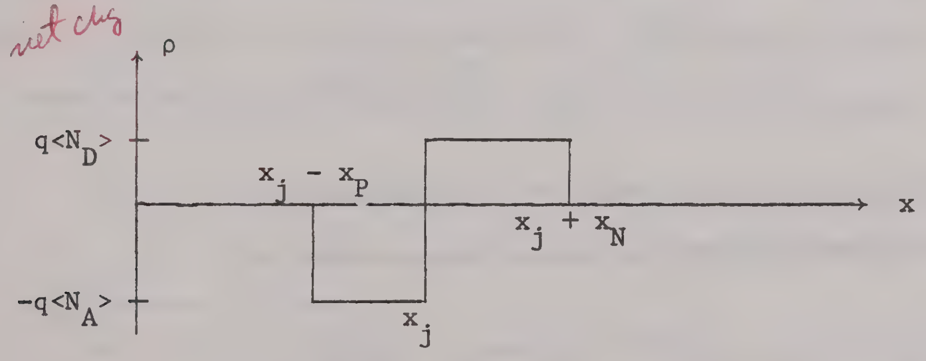
Note that the energy diagram suggests that we have three regions:

I. $0 < x < x_j - x_P$ $\quad E = 0 \quad$ neutral ($\rho=0$)

II. $x_j - x_P < x < x_j + x_N$ $\quad E \neq 0$

III. $x_j + x_N < x < w$ $\quad E = 0 \quad$ neutral ($\rho=0$)

We call the two neutral regions the bulk. The middle region is called the space charge or depletion region. Its origin resides in the fact that when contact between the two pieces of silicon is established a hole-electron concentration gradient exists. Thus, holes diffuse into the N-region, electrons into the p-region. The N-material ~~assures~~ assumes a positive charge, the p-material a negative charge. A field results which forces $J_n = 0$ and $J_p = 0$, i.e., is demanded by the equilibrium assumption. In this field region electrons and holes; mobile carriers; do not exist. They are essentially swept out. The fixed charges, however, remain. The uncompensated charge looks as follows:



Uncompensated Fixed Charge

The quantities x_P and x_N are not known. We may estimate them as follows: Charge and field are related by Gauss~~i~~'s Law. Hence, if $y = x - x_j$

$$\varepsilon \frac{d\varepsilon_P}{dy} = -q\langle N_A\rangle \qquad\qquad \varepsilon \frac{d\varepsilon_N}{dy} = q\langle N_D\rangle$$

Integrate from $y = -x_P$ to y $\qquad\qquad$ $y = y$ to x_N

$$E_P(y) - E_P(-x_P) = -\frac{q}{\varepsilon}\langle N_A\rangle (y+x_P) \qquad\qquad E_N(x_N) - E_N(y) = \frac{q}{\varepsilon}\langle N_D\rangle (x_N - y)$$

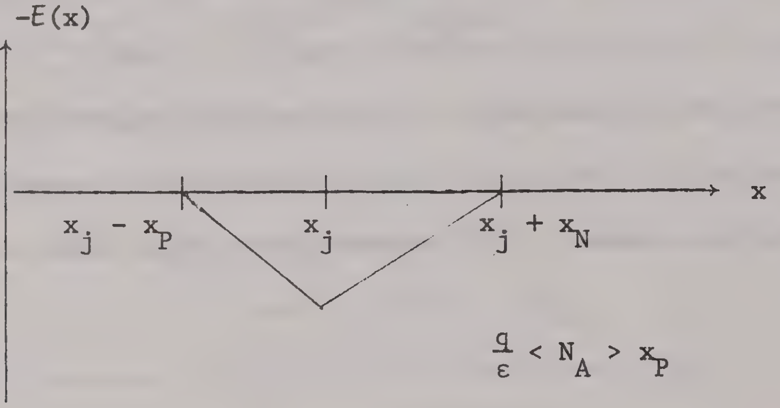
But the space charge region borders on the neutral regions. In the neutral regions $J_N = 0$, $J_P = 0$ demands $\mathcal{E} = 0$. Hence, $\mathcal{E}_P(-x_P) = \mathcal{E}_N(x_N) = 0$. Then

$$\frac{dv_P}{dx} = \frac{q}{\varepsilon} < N_A > (y+x_P) \qquad \frac{dv_N}{dx} = -\frac{q}{\varepsilon} < N_D > (y-x_N)$$

Since the field is zero at the two edges of the space charge region the total charge enclosed in the space charge region must be zero. This is equivalent to the statement $\mathcal{E}_P(0) = \mathcal{E}_N(0)$ or

$$(1) \quad < N_A > x_P = < N_D > x_N$$

The field looks as follows.



Note that the field points in the negative x-direction.

The potential drop across the space charge region is:

$$\Delta v = [v_P(y{=}-x_P) - v_P(0)] - [v_N(y{=}x_N) - v_N(0)]$$

$$= -\frac{q}{\varepsilon} < N_A > \int_{-x_P}^{0} (y+x_P)\,dy - \frac{q}{\varepsilon} < N_D > \int_{0}^{x_N} (x_N-y)\,dy$$

Hence

$$-\Delta v = \frac{q}{\varepsilon} < N_A > \left[-\frac{x_P^2}{2} + x_P^2 \right] + \frac{q}{\varepsilon} < N_D > \left[x_N^2 - \frac{x_N^2}{2} \right]$$

or

$$(2) \quad -\Delta v = \frac{q<N_A>}{2\varepsilon} x^2 + \frac{q<N_D>}{2\varepsilon} x_N^2$$

Note that according to the energy diagram $-\Delta v$ is a positive quantity: The diode assumes a natural reverse bias. (N is positive with respect to P.)

Equations (1) and (2) may be combined to express x_N and x_P in terms of the potential difference Δv.

$$-\Delta v = \frac{q}{2\varepsilon}\left[< N_A > \left(\frac{<N_D>}{<N_A>}\right)^2 + < N_D > \right] x_N{}^2$$

$$= \frac{q}{2\varepsilon}\,\frac{<N_D>}{<N_A>}\left(< N_D > + < N_A >\right) x_N{}^2$$

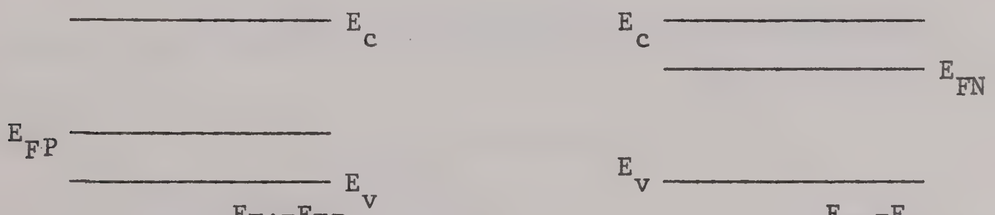
or

$$(3) \qquad x_N = \left(\frac{2\varepsilon(-\Delta v)}{q}\right)^{1/2}\left(\frac{<N_A>}{<N_D>\left(<N_D>+<N_A>\right)}\right)^{1/2}$$

The expression for x_p may be obtained by subscript exchange A → D, D → A. Then

$$(4) \qquad x_P = \left(\frac{2\varepsilon(-\Delta v)}{q}\right)^{1/2}\left(\frac{<N_D>}{<N_A>\left(<N_D>+<N_A>\right)}\right)^{1/2}$$

The next quantity which we have to evaluate is Δv, the potential difference from bulk to bulk or across the space charge region. In order to do this consider that we break our junction exactly at x_j and form two pieces of uniformly doped, separated material. Then we have the following situation:



$$<N_A> = n_i \varepsilon^{\frac{E_{Fi}-E_{FP}}{kT}} \qquad <N_D> = n_i \varepsilon^{\frac{E_{FN}-E_{Fi}}{kT}}$$

$$E_{FP} = E_{Fi} - kT \ln \frac{<N_A>}{n_i} \qquad E_{FN} = E_{Fi} + kT \ln \frac{<N_D>}{n_i}$$

If we make contact the relative shift of E_c or E_v or E_{Fi} will be the difference in the two fermi levels. Since

$$qE = \frac{dE_c}{dx} = -q\,\frac{dv}{dx}$$

$$dE_e = -q\ dv$$

$$E_c(x_N) - E_c(-x_p) = -q[v(x_N) - v(-x_p)]$$

$$= q\Delta v$$

or $$q\Delta v = E_{FP} - E_{FN}$$

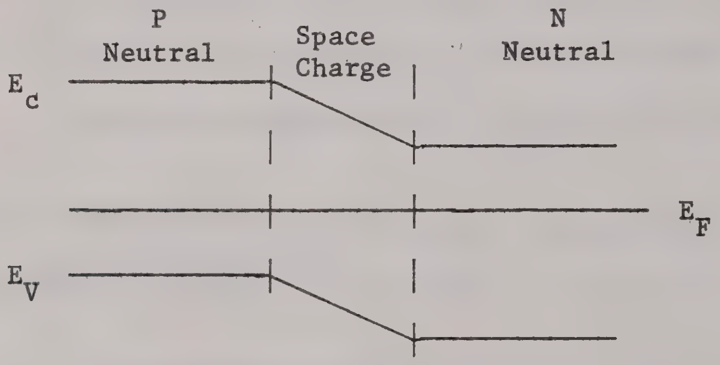
$$= -\ kT\ \ln \frac{<N_D><N_A>}{n_i^2}$$

Hence

$$(5) \quad -\Delta v = \frac{kT}{q} \ln \frac{<N_D><N_A>}{n_i^2}$$

If we substitute (5) into (3) or (4) the equilibrium depletion layer dimensions x_N and x_P can be calculated.

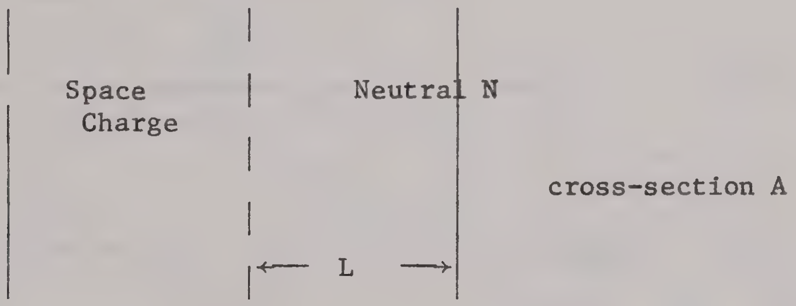
Let us now consider the junction in some more detail.



Recall that holes behave like bubbles and electrons like balls. This suggests that there is a hole flow from N → P and an electron flow from P → N. The source of these two currents are minority carriers. They become majority carriers after injection. The process is called majority carrier injection. Let's consider the source for the holes on the N-side. They come from the neutral region in which we have no field. Thus in the neutral region they must exist as a diffusion current. But we have the phenomena of recombination and generation which would suggest that the average life of a pair is τ. On the N-side

$$n_0 \approx \langle N_D \rangle >> p_0 = \frac{n_i^2}{\langle N_D \rangle}$$

If we assume that we remove all holes next to the space charge region the effect on the majority carrier n_0 will be very small. New holes will move into the vacated area. Let's assume for a second that they come from the N-bulk. Since they have a finite velocity it takes time. If this time is longer than τ the holes will recombine and cannot be used to make up our deficiency. But this says that there is a finite number of minority carriers available for injection. Let's estimate this number as follows:



If all holes in the volume LA recombine in time τ the recombination current would be

$$\overset{I_R}{J_R} = \frac{1}{\tau}\int_0^L q\, P_{n0} A\, dx = \frac{q\, P_{n0}}{\tau} AL$$

The compensating hole current which flows into that area from the right would be roughly given by

$$I_p = AqD_p \frac{DP}{dx} \approx qD_p A \frac{P_{n0}}{L}$$

but the two currents are equal. Hence

$$\frac{D_p}{L} = \frac{L}{\tau}$$

or

$$L = \sqrt{D_p \tau}$$

Hence

$$I_p = \frac{qP_{n0} D_p}{\sqrt{D_p \tau}} A = qP_{n0} \sqrt{\frac{D_p}{\tau}} A$$

We call L the diffusion length and note that the effective diffusion velocity is $\nu = \sqrt{\frac{D_p}{\tau}} = \frac{L}{\tau}$.

Let us now look at the contact potential a little closer. We have

$$-\Delta v = \frac{kT}{q} \ln \frac{\langle N_D \rangle \langle N_A \rangle}{n_i^2}$$

but

$$n_{n0} = \langle N_D \rangle \qquad \text{and} \qquad n_{p0} = \frac{\langle N_A \rangle}{n_i^2}$$

Hence

$$(1) \qquad n_{p0} = n_{n0} \varepsilon^{-\frac{q(-\Delta v)}{kT}}$$

Recall our ideal gas. The gas molecule density is of the form

$$n_G = A\varepsilon^{-\frac{E}{kT}}$$

if

$$n_{G1} = A\varepsilon^{-\frac{E_1}{kT}}$$

we write

$$(2) \qquad n_G = n_{GI} \varepsilon^{-\frac{E-E_1}{kT}}$$

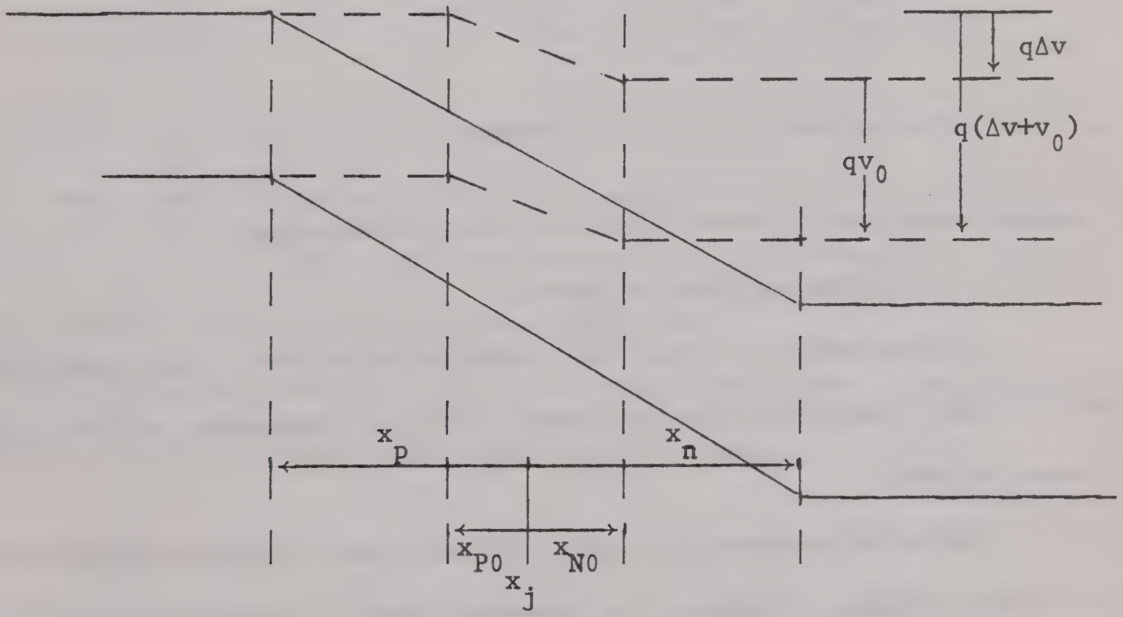
But if we recall the definition of potential we may write Eq. (1) as

$$(1a) \qquad n(P) = n(N) \varepsilon^{-\frac{E_c(P)-E_c(N)}{kT}}$$

The contact potential is only a statement of Boltzmann equilibrium between two levels which are separated by $E_c(P) - E_c(N) = E_v(P) - E_v(N) = -q\Delta v$. But this implies that for our equilibrium situation the minority carriers which are injected are immediately replaced by majority carriers which are energetic enough to cross the potential barrier. This flow, however, can be cut-off completely by increasing the potential barrier by external means. The result will be that we will experience a net current density

$$J = \frac{qp_{n0}D_p}{L_P} + \frac{qn_{p0}D_n}{L_N}$$

Since this current flows by diffusion it will not cause a potential drop in the neutral region close to the junction. Since the source of the current is minority carrier based we might expect it to be very small. The resulting energy diagram is as shown:



Note that $x_{P0} \to x_P$, $x_{N0} \to x_N$. The width of the depletion region is increased in particular, the uncompensated donor charge becomes

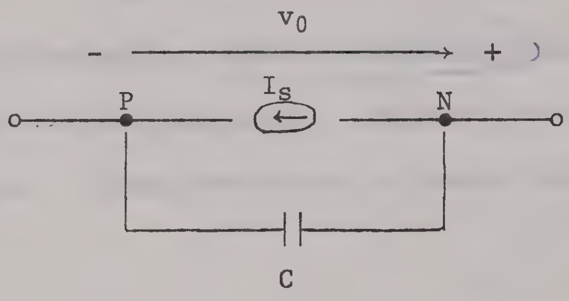
$$Q = q\langle N_D\rangle x_N$$

where $$x_N = \left(\frac{2\varepsilon(v_0+|\Delta v|)}{q}\right)^{1/2} \left(\frac{<N_A>}{<N_D>(<N_D>+<N_A>)}\right)^{1/2}$$

If we change v_0 we change Q. The result is a differential capacitance which we define by

$$C = \frac{dQ}{dv_0} = \left(\frac{q\varepsilon}{2(v_0+|\Delta v|)}\right)^{1/2} \left(\frac{<N_A><N_D>}{<N_D> + <N_A>}\right)^{1/2} \quad \text{farads/m}^2$$

Here $v_0 + |\Delta v|$ is the total reverse bias across the junction: N positive with respect to P. Note that the differential capacitance is voltage dependent. The equivalent circuit for the reverse biased junction becomes simply



$$I_s = \left(\frac{qp_{n0}D_p}{L_P} + \frac{qn_{p0}D_n}{L_N}\right)A \text{ amps}$$

$$C = c(v=0)A\left[\frac{|\Delta v|}{v_0 + |\Delta v|}\right]^{1/2} \text{ farads}$$

The assumptions that we have made here are the following:

1. The doping profile is abrupt.

2. The potential barrier was raised so high by v_0 that hole injection form P → N and electron injection from N → P; i.e., minority carrier injection, no longer occurs.

3. Majority carrier injection: Holes from N → P and electrons from P → N occurs readily. However, the minority carrier concentration is low and occurs only from a region next to the junction due to lifetime and velocity considerations. The resulting current is small.

4. The carrier concentration in the space charge region is very small, less than n_i. The space charge region is a depletion region.

5. Since the current is small and a large potential is assumed to satisfy 2 the resistive voltage drop in the neutral bulk is neglected. Hence v_0 is the applied voltage.

In view of these assumptions it is clear that our equivalent circuit is only roughly true. In addition the restriction to an abrupt profile is not very desirable. Let us attempt some inprovement.

Consider first of all the space charge region. Since it is a piece of semiconductor it exhibits generation-recombination transitions. But in this region p and n are nearly zero. Thus recombination is essentially non-existent. But the thermal generation rate is fixed. Hence

$$G = r_{vc} N_v N_c = r_{cv} n_i^2$$
$$= \frac{n_i^2}{\tau(n_0+p_0)}$$

is the thermal generation rate of carriers in the space-charge region. These pairs are immediately swept out of this region by the high field. The holes are inserted into the P-region and the electrons into the N-region. Thus we are again talking about majority carrier injection.

A more accurate analysis of lifetime in a crystal with imperfections indicates that the thermal generation rate takes the form

$$G = \frac{n_i}{2\tau}$$

The current which is associated with this mechanism is

$$J = \int_{-x_P}^{x_N} qG\, dy$$
$$= q\,\frac{n_i}{2\tau}\,(x_N+x_P)$$

But for the abrupt junction this becomes

$$J = q\,\frac{n_i}{2\tau}\left[\frac{2\varepsilon(v_0+|\Delta v|)}{q}\right]^{1/2}\left[\frac{\frac{<N_A>}{<N_D>}+\frac{<N_D>}{<N_A>}}{<N_D>+<N_A>}\right]^{1/2}$$

The current which originates in the bulk is of the form

$$J_s = \frac{qp_{n0}D_p}{L_P} + \frac{qn_{p0}D_n}{L_N}$$
$$= \frac{qp_{n0}L_P}{\tau_P} + \frac{qn_{p0}L_N}{\tau_N}$$

The space-charge generated term is

$$J = \frac{qn_iw}{2\tau} \qquad w = x_N + x_P$$

Hence

$$\frac{J_s}{J} = 2\,\frac{p_{n0}L_P}{\tau_P}\,\frac{\tau}{n_iw} + 2\,\frac{n_{p0}L_N}{\tau_N}\,\frac{\tau}{n_iw}$$
$$= \frac{2n_i}{<N_D>}\,\frac{L_P}{w}\,\frac{\tau}{\tau_P} + \frac{2n_i}{<N_A>}\,\frac{L_N}{w}\,\frac{\tau}{\tau_N}$$

Normally τ and τ_P and τ_N are of the same order of magnitude. L_P and L_N are larger than w. But $n_i << <N_D>$, $n_i << <N_A>$. This would suggest that $J_s/J << 1$ or $J >> J_s$. The space charge generated current would dominate the reverse behavior. We must therefore add this current generator to our equivalent circuit.

Hence

$$I_{REV} = \left(\frac{qp_{n0}}{L_P}D_p + \frac{qn_{p0}}{L_N}D_n\right)A + \frac{qn_i}{2\tau}\left[\frac{2\varepsilon(v_0+|\Delta v|)}{q}\right]^{1/2}\left[\frac{\frac{<N_A>}{<N_D>}+\frac{<N_D>}{<N_A>}}{N_A + N_D}\right]^{1/2} A$$

$$= I_{BULK} + I_{SPACE\text{-}CHARGE}$$

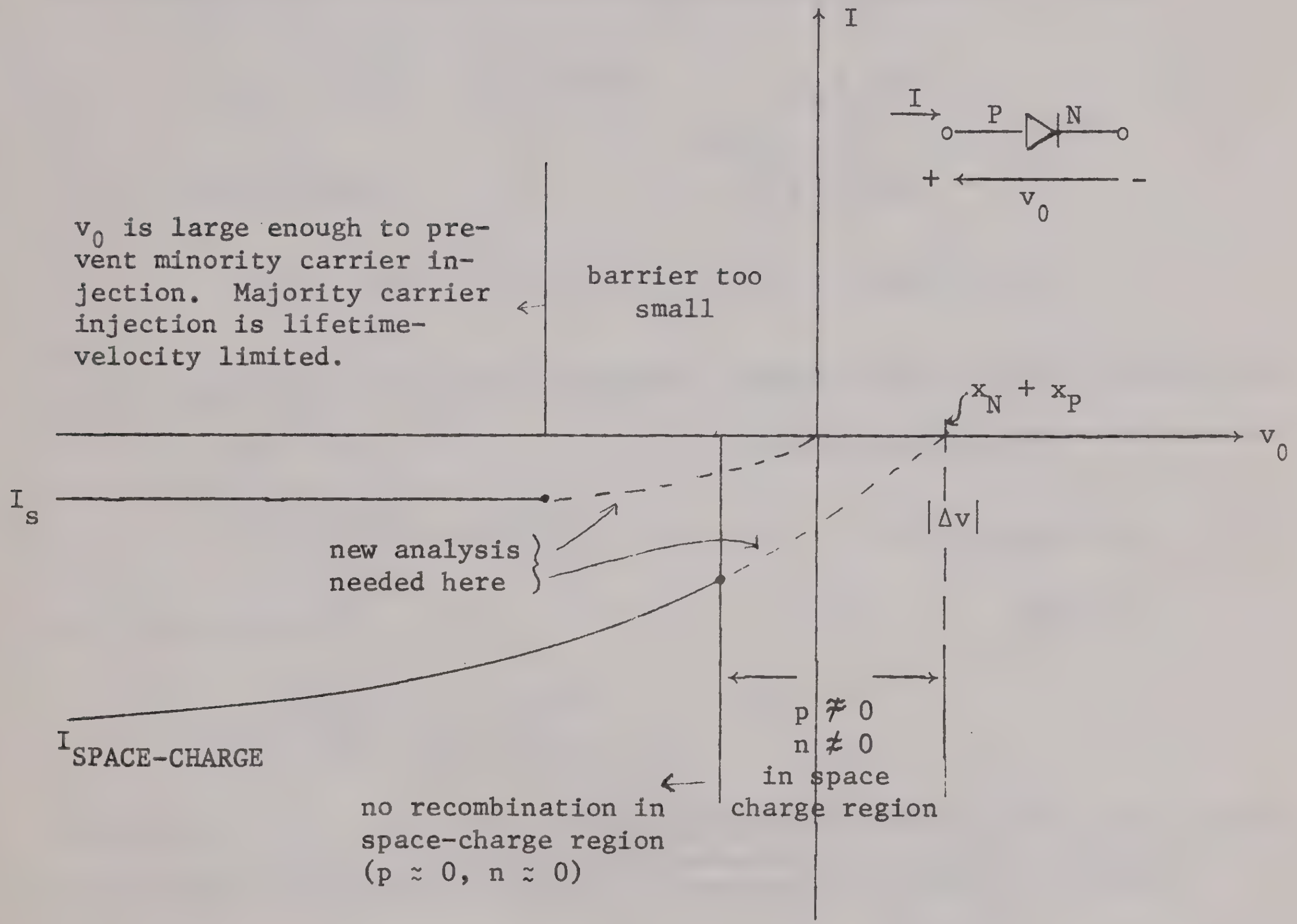
The v-I characteristic of the diode is shown on page 77. The small signal equivalent is obtained as follows

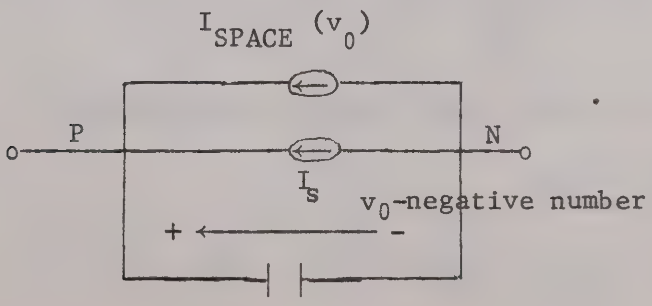
$$I_{REV}(v_0+\Delta v_0) = I_{REV}(v_0) + \left.\frac{dI_{REV}}{dv_0}\right|_{v_0} \Delta v_0$$

$$\frac{\Delta I_{REV}}{\Delta v_0} = \frac{I_{REV}(v_0+\Delta v_0) - I_{REV}(v_0)}{\Delta v_0} = \left.\frac{dI_{REV}}{dv_0}\right|_{v_0} \equiv q_R = \frac{1}{r_R}$$

$$r_R = 2\,\frac{v_0 + |\Delta v|}{I_{REV}}$$

Typical values for a switching diode biased to -20 volts or so are $I_{REV} = 1\ \mu A$. This yields $r_R \simeq 20\times10^6$ ohms. A differential capacitance of 2 pf or so is typical.





Let us now return to our second assumption: Uniform doping. The doping profile has the form

$$N_A = N_d \varepsilon^{-(ax)^2} - N_B$$

in the vicinity of the junction:

$$N_A(x_j + \Delta x) = N_A(x_j) + \left.\frac{dN_A}{dx}\right|_{x_j} \Delta x$$

$$= -2a^2 x_j N_j \varepsilon^{-(ax_j)^2} \Delta x$$

$$= -2a^2 x_j N_B \Delta x$$

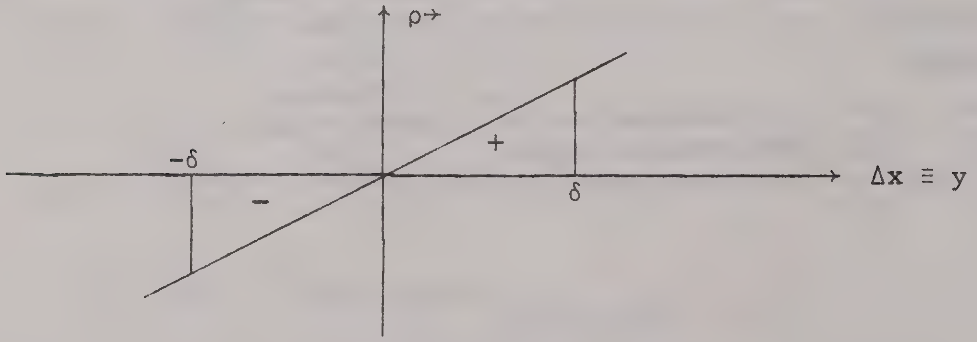
$$= -2aN_B \left[\ln \frac{N_s}{N_B}\right]^{1/2} \Delta x$$

$$\equiv -\alpha \Delta x$$

If we assume that all mobile charges are removed from the neighborhood of the junction the charge density will be

$$\rho = +qN_A(x_j + x)$$

$$= q\alpha\Delta x$$



Then Poisson's equation for the space charge region becomes

$$\frac{dE}{dy} = \frac{q}{\varepsilon} dy \qquad -\delta \le y \le \delta$$

$$E(y) - E(-\delta) = \frac{q}{\varepsilon}\frac{\alpha}{2}(y^2 - \delta^2)$$

Since the doping is non-uniform the quasi-neutral region will have a field

$$\mathcal{E}_0 = \frac{kT}{q}\frac{1}{N_A}\frac{dN_A}{dx} = -\frac{kT}{q}\frac{2a^2x\,N_s\varepsilon^{-(ax)^2}}{N_s\varepsilon^{-(ax)^2} - N_B}$$

In the neighborhood of the junction this may be approximated by using the results of page 78. It yields

$$\mathcal{E}_0 \approx \frac{kT}{q}\frac{1}{|\Delta x|} = \frac{kT}{q}\frac{1}{|y|}$$

The field $\mathcal{E}_0$ and the space charge field $\mathcal{E}$ are equal at $x = -\delta$. Hence

$$\mathcal{E}(y) = \mathcal{E}_0(-\delta) + \frac{q}{\varepsilon}\frac{\alpha}{2}(y^2-\delta^2)$$

The potential becomes

$$-[v(\delta) - v(-\delta)] = 2\mathcal{E}_0(\pm\delta)\delta + \frac{q}{\varepsilon}\frac{\alpha}{2}\left(\frac{\delta^3}{3} - \delta^3 + \frac{\delta^3}{3} - \delta^3\right)$$

or

$$\begin{aligned} v_{NP} &= -v_{PN} \\ &= \frac{2}{3}\frac{q\alpha\delta^3}{\varepsilon} - 2\delta\frac{kT}{q}\frac{1}{\delta} \end{aligned}$$

or, solving for δ:

$$2\delta = \left[\frac{12\varepsilon}{9\alpha}\left(v_{NP}+2\frac{kT}{q}\right)\right]^{1/3} = \omega$$

The potential v_{NP} is the total reverse potential across the depletion region. It includes the built-in potential. In order to evaluate it we write

$$\begin{aligned} v_{NP0} &= \frac{kT}{q}\ln\frac{p_{p0}(-\delta_0)}{p_{n0}(\delta_0)} = \frac{kT}{q}\ln\frac{p_{p0}(-\delta_0)n_{n0}(\delta_0)}{n_i^2} \\ &= 2\frac{kT}{q}\ln\frac{\alpha\delta_0}{n_i} \end{aligned}$$

but

$$v_{NP0} = \frac{2}{3}\frac{q\alpha\delta_0^3}{\varepsilon} - 2\frac{kT}{q} = 2\frac{kT}{q}\ln\frac{\alpha\delta_0}{n_i}$$

In order to find δ_0 then we solve a transcendental equation first. We then evaluate

$$v_0 = v_{NP0} + 2\,\frac{kT}{q}$$

and write

$$\omega = \left(\frac{12\varepsilon}{9\alpha}(v_{NP0}+v_0)\right)^{1/3}$$

The amount of charge which we have in one side of the depletion region is

$$\begin{aligned} Q &= q\,\frac{1}{2}\,\delta\;\alpha\delta \\ &= \frac{1}{8}\,q\,\alpha\left(\frac{12\varepsilon}{9\alpha}(v_{NP0}+v_0)\right)^{2/3} \end{aligned}$$

then

$$\begin{aligned} C &\equiv \frac{dQ}{dv_{NP}} \\ &= \left[\frac{q\alpha\varepsilon^2}{12}\,\frac{1}{v_{NP0}+v_0}\right]^{1/3} \end{aligned}$$

The depletion layer capacitance varies a $v^{-1/3}$ for the linear grade. Our earlier results indicate that for an abrupt junction the capacitance varies as $v^{-1/2}$. In real life one measures $C \sim V^{-1/m}$ $\quad 2 < m < 3$.

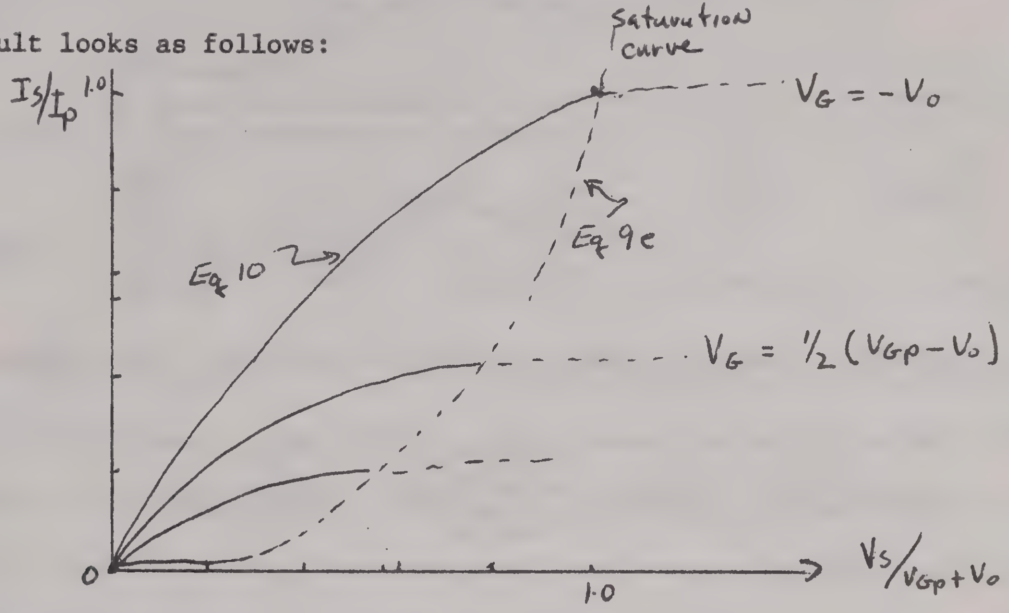
region of the JFET. It is evaluated as follows:

(a) assume some value of V_G such that $-V_0 < V_G < V_{GP}$

(b) calculate $V_{SS} = V_{GP} - V_G$

(c) Evaluate (11) for fixed V_{SS} and varying V_S.

The result looks as follows:



Note that we have extended the characteristics into the pinched-off region. The gate has lost control over the channel. The N^+-contact is surrounded by a depletion region and behaves no longer like an ohmic contact. The current becomes independent of V_S.

SMALL SIGNAL EQUIVALENT

The description of the JFET in terms of Eq. (9b) is of the form:

$$I_S = I_S(V_S, V_G)$$

Hence

$$I_S + \Delta I_S = I_S(V_S, V_G) + \frac{\partial I_S}{\partial V_S} \Delta V_S + \frac{\partial I_S}{\partial V_G} \Delta V_G$$

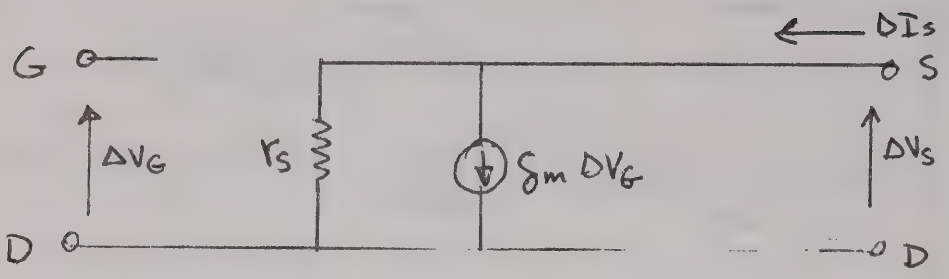
or

$$\Delta I_S = \frac{\partial I_S}{\partial V_S} \Delta V_S + \frac{\partial I_S}{\partial V_G} \Delta V_G \tag{12}$$

We define

$$(13) \qquad \frac{1}{r_S} = \frac{\partial I_S}{\partial V_S} \quad \text{and } q_m = \frac{\partial I_S}{\partial V_G}$$

The equivalent circuit for the device becomes:



We may evaluate r_S and q_m by using Eq. (9b) or Eq. (11). This yields:

$$\frac{1}{3I_P}\frac{\partial I_S}{\partial V_S} = \frac{1}{V_{GP}+V_0} - \frac{(V_S+V_G+V_0)^{1/2}}{(V_{GP}+V_0)^{3/2}} = \frac{R_0}{V_{GP}+V_0} \quad \frac{\partial I_S}{\partial V_S}$$

Hence

$$(14) \qquad \frac{\partial I_S}{\partial V_S} = \frac{1}{R_0} \left(1-\left(\frac{V_S+V_G+V_0}{V_{GP}+V_0}\right)^{\frac{1}{2}}\right)$$

$$(14a) \qquad r_S = \frac{R_0}{1 - \left(\frac{V_S+V_G+V_0}{V_{GP}+V_0}\right)^{1/2}}$$

The saturation condition is

$$V_S = V_{SS} = V_{GP}-V_G$$

Examination of (14a) yields the conclusion that $r_S \to \infty$ at saturation. This of course is reasonable in view of our discussion of the saturation condition. If $V_S = 0$ and $V_G = -V_0$ $r_S = R_0$. This again is reasonable in view of our discussions with $I_S = 0$.

The transconductance may be computed from (9b)

$$\frac{R_0}{V_{GP}+V_0}\,\frac{\partial I_S}{\partial V_G} = \frac{1}{V_{GP}+V_0} - \frac{(V_S+V_G+V_0)^{\frac{1}{2}}}{(V_{GP}+V_0)^{3/2}} - \frac{1}{V_{GP}+V_0} + \frac{(V_G+V_0)^{\frac{1}{2}}}{(V_{GP}+V_0)^{3/2}}$$

$$(15) \qquad \frac{\partial I_S}{\partial V_G} = \frac{1}{R_0}\left[\left(\frac{V_G + V_0}{V_{GP}+V_0}\right)^{\frac{1}{2}} - \left(\frac{V_S+V_G+V_0}{V_{GP}+V_0}\right)^{\frac{1}{2}}\right]$$

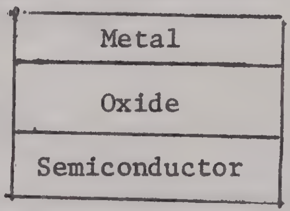
At saturation the transconductance becomes

$$(15a) \qquad \frac{\partial I_S}{\partial V_G} = \frac{1}{R_0}\left[\left(\frac{V_G+V_0}{V_{GP}+V_0}\right)^{\frac{1}{2}} - 1\right]$$

If $V_G = - V_0$ the g_m is $- \frac{1}{R_0}$. The minus sign indicates phase inversion. This value is the maximum g_m which occurs. The channel is near pinch-off at the source and near forward bias at the drain. If we stay at saturation but increase the gate bias the g_m decreases and goes to zero as the channel pinches off a~~s~~t both source and drain.

METAL-OXIDE SEMICONDUCTOR DEVICES

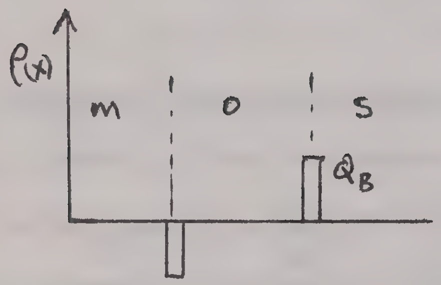
The oxidation of silicon is an integral part of integrated circuit manufacturing. It yields silicon dioxide; SiO_2; which is also known as quartz. Since this material is a very poor conductor or good insulator it would appear that it is ideally suited for capacitor manufacturing. The resulting structure would appear as follows:

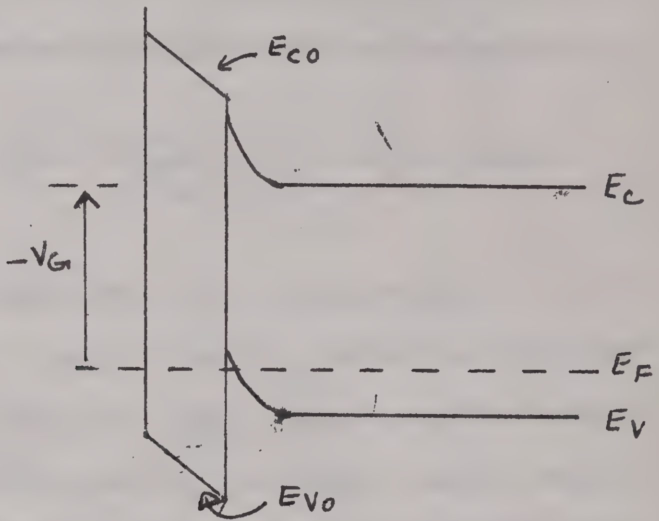


Let us now assume that the semiconductor is P-type and that a voltage V_G is applied across the device:

(a) Assume V_G is negative

This would imply that some holes or majority carriers would be swept up to the $S_i - S_iO_2$ interface. These holes form a positive excess charge at the interface which is balanced by an equal but opposite charge on the metal. The charge density and energy diagram is as shown:

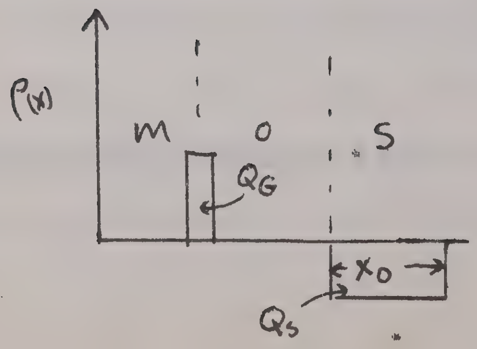


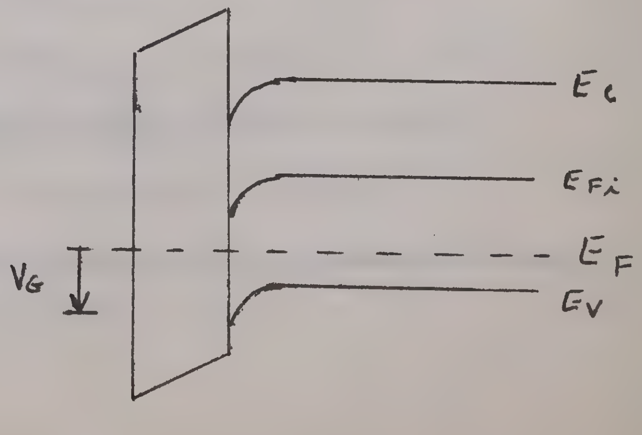


here we have assumed that the S_iO_2 is a wideband semiconductor, $\Delta E \sim 8eV$. The field in the oxide is in the negative x-direction. Note that the movement of holes towards the surface causes an accumulation layer, i.e. a region in which $P_P > P_{PO} = N_A$.

(b) Assume V_G is positive.

If V_G is positive holes will be repelled from the surface or interface. If V_G is not too large a depletion region forms since uncompensated acceptors are present. The charge configeration is as follows:



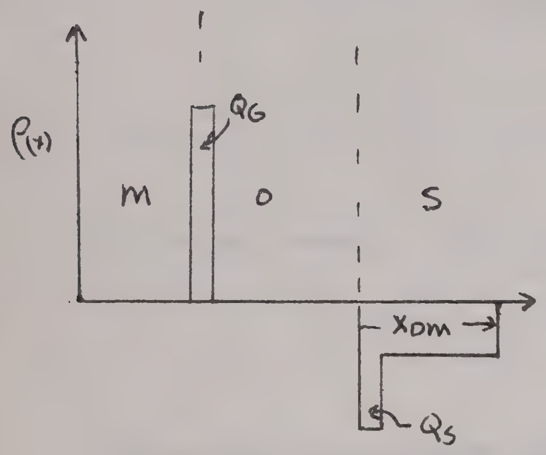


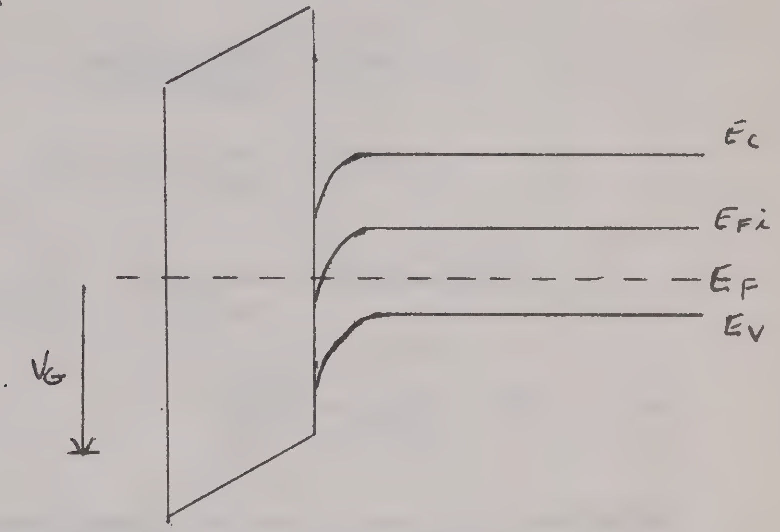
The charge in the depletion region is

$$Q_S = -qx_D N_A$$

The surface of the semiconductor is depleted of majority carriers. The carrier concentration near the surface is $P_{P0} > P_P(0) > n_i$.

Suppose now that we make the metal even more positive this would result in the attraction of electrons to the surface. They in turn cause a negative charge to appear at the interface.





Note that the energy diagram suggests that $n_p > n_i$ near the surface. We spear of inversion: the accumulation of minority carriers. As soon as we form an inverted region the depletion layer will essentially remain constant with further voltage increases. We therefore have a maximum depletion layer width. The semiconductor charge becomes

$$Q_S = -Q_n - qN_A x_{DMAX}$$

We may analyse the positive potential case in a straight forward manner. In the depletion mode we have

$$\varepsilon_{Si}\,\frac{dE}{dx} = -q\,N_A \qquad 0 \leq x < x_D$$

$$dE = -\,\frac{q}{\varepsilon_{Si}}\,N_A\,dx$$

$$E(x_D)-E(x) = -\,\frac{q}{\varepsilon_{Si}}\,N_A(x_D-x)$$

But $$E(x_D) = 0$$

$$E(x) = \cancel{-}\,\frac{q}{\varepsilon_{Si}}\,N_A(x_D-x)$$

$$dV = -\,\frac{q}{\varepsilon_{Si}}\,N_A(x_D-x)dx$$

$$V(x_D)-V(0) = -\,\frac{q}{\varepsilon_{Si}}\,N_A(x_D^2 - \frac{x_D^2}{2})$$

or $$V(0) - V(x_D) = \frac{q\,N_A x_D^2}{2\varepsilon_{Si}}$$

We define this quantity as the surface potential. Note that it is given by

$$V_S = -\,\frac{E_c(0)-E_c(\infty)}{q} = -\,\frac{E_{Fi}(0)-E_{Fi}(\infty)}{q}$$

It may be used to write

$$dV = -2\,\frac{V_S}{x_D^2}\,(x_D-x)\;dx$$

or

$$V(x_D)-V(x) = -2\,\frac{V_S}{x_D^2}\,(x_D^2 - \frac{x_D^2}{2} - xx_D + \frac{x^2}{2})$$

$$V(x) - V(x_D) = \frac{V_S}{x_D^2}\,(x_D^2 - 2xx_D + x^2)$$

$$V(x) - V(x_D) = V_S\,(1 - \frac{x}{x_D})^2$$

If all potentials are referenced to the potential in the neutral bulk we get

$$V(x) = V_S\left(1 - \frac{x}{x_D}\right)^2$$

? why

The onset of strong inversion is defined by

$$n_S = n(0) = N_A = P_{P0}$$

In the neutral bulk we have

$$n = n_i\,\varepsilon^{\frac{E_F - E_{Fi}(\infty)}{kt}} \qquad P = N_A = n_i\varepsilon^{-\frac{E_F - E_{Fi}(\infty)}{kt}}$$

Hence $$kt\,\ell n\frac{N_A}{n_i} = E_{Fi}(\infty) - E_F$$

The strong inversion condition yields

$$kt\,\ell n\frac{N_A}{n_i} = E_F - E_{Fi}(0)$$

Hence, the onset of inversion is given by

$$E_F - E_{Fi}(0) = E_{Fi}(\infty) - E_F$$

$$2E_F = E_{Fi}(\infty) + E_{Fi}(0) = \cancel{2kt\,\ell n\frac{N_A}{n_i}}$$

Since $qV_{Si} = E_{Fi}(\infty) - E_{Fi}($

and $E_{Fi}(\infty) - E_{Fi}(0) = 2kT\ln$

~~The inversion condition becomes therefore~~

$$V_{Si} = \cancel{2\frac{E_F}{q} = 2E_F} + 2\frac{kt}{q}\,\ell n\frac{N_A}{n_i}$$

The width of the depletion region is

$$x_D = \left[\frac{2\varepsilon_{Si}}{qN_A}V_S\right]^{1/2}$$

The maximum depletion layer width occurs when $V_S = V_{Si}$. Hence

$$x_{DMAX} = \left[\frac{2\varepsilon_{Si}}{qN_A} V_{Si}\right]^{1/2}$$

If $V_S > V_{Si}$ the charge in the semiconductor becomes

$$Q_s = Q_n + Q_B$$

where $Q_B \equiv -qN_A\, x_{DMAX}$

Q_n: inversion charge.

In order to calculate the capacitance for our structure let us write:

$$(1) \qquad V_G = V_{ox} + V_S$$

Here V_{ox} is the potential drop across the oxide. If the dielectric is perfect, i.e. if no charges are present in the dielectric, the oxide field is constant

$$(2) \qquad E_{ox} = \frac{V_{ox}}{x_{ox}}$$

in the semiconductor we have

$$\varepsilon_{Si} \frac{dE_S}{dx} = \rho(x)$$

$$E_S(\infty) - E_S(0) = \frac{1}{\varepsilon} Q_S$$

or $$(3) \qquad E_S(0) = -\frac{Q_S}{\varepsilon_{Si}}$$

But at the interface

$$(4) \qquad \varepsilon_{Si} E_S = \varepsilon_{ox} E_{ox} \quad \text{at } x = 0$$

Hence

$$V_{ox} = x_{ox} E_{ox} = x_{ox} \frac{\varepsilon_{Si}}{\varepsilon_{ox}} E_S(0) = -\frac{x_{ox}}{\varepsilon_{ox}} Q_S$$

But the oxide capacitance/unit area is

$$C_{ox} = \frac{\varepsilon_{ox}}{x_{ox}}$$

Hence

$$(5) \qquad V_{ox} = -\frac{Q_S}{C_{ox}}$$

If (5) is substituted into Eq. (1) we have

$$(1a) \qquad V_G = V_S - \frac{Q_S}{C_{ox}}$$

We now define the differential capacitance of the structure by

$$(6) \qquad C = -\frac{dQ_S}{dV_G}$$

Then

$$(1b) \qquad V_S = V_G + \frac{1}{C_{ox}} Q_S$$

or

$$\frac{dV_S}{dQ_S} = \frac{dV_G}{dQ_S} + \frac{1}{C_{ox}}$$

But

$$\frac{1}{C_S} = -\frac{dV_S}{dQ_S}$$

Hence

$$(7) \qquad \frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_S}$$

or

$$(7a) \qquad C = \frac{C_{ox} C_S}{C_S + C_{ox}}$$

The total capacitance is therefore the series capacitance of the oxide and the semiconductor. We may write

$$(7b) \qquad \frac{C}{C_{ox}} = \frac{1}{1 + \frac{C_{ox}}{C_S}}$$

In order to understand the behavior of the capacitance in the depletion region: $V_G > 0$ But $V_G >> 0$, i.e. moderate forward voltage, consider

$$(1a) \qquad V_G = -\frac{Q_S}{C_0} + V_S$$

But

$$Q_S = -qN_A x_D = -qN_A \left(\frac{2\varepsilon_{Si}}{qN_A} V_S\right)^{\frac{1}{2}}$$

Hence

$$(1b) \qquad V_G = V_S + qN_A \frac{x_{ox}}{\varepsilon_{ox}} \left(\frac{2\varepsilon_{Si}}{qN_A} V_S\right)^{\frac{1}{2}}$$

If we solve this equation for $\sqrt{V_S}$ we find

$$(8) \quad \sqrt{V_S} = -\frac{qN_A}{2}\frac{x_{ox}}{\varepsilon_{ox}}\left(\frac{2\varepsilon_{Si}}{qN_A}\right)^{\frac{1}{2}} + \left[V_G + \left(\frac{qN_A}{2}\frac{x_{ox}}{\varepsilon_{ox}}\left(\frac{2\varepsilon_{Si}}{qN_A}\right)^{\frac{1}{2}}\right)^2\right]^{\frac{1}{2}}$$

We now write Eq. (7b) as

$$(7b) \qquad \frac{C_{ox}}{C} = 1 + \frac{C_{ox}}{C_S} = 1 + \frac{\varepsilon_{ox}}{x_{ox}}\frac{x_D}{\varepsilon_{Si}}$$

$$= 1 + \frac{\varepsilon_{ox}}{x_{ox}}\frac{1}{\varepsilon_{Si}}\left(\frac{2\varepsilon_{Si}}{qN_A} V_S\right)^{\frac{1}{2}}$$

Substitution of (8) into (7b) yields

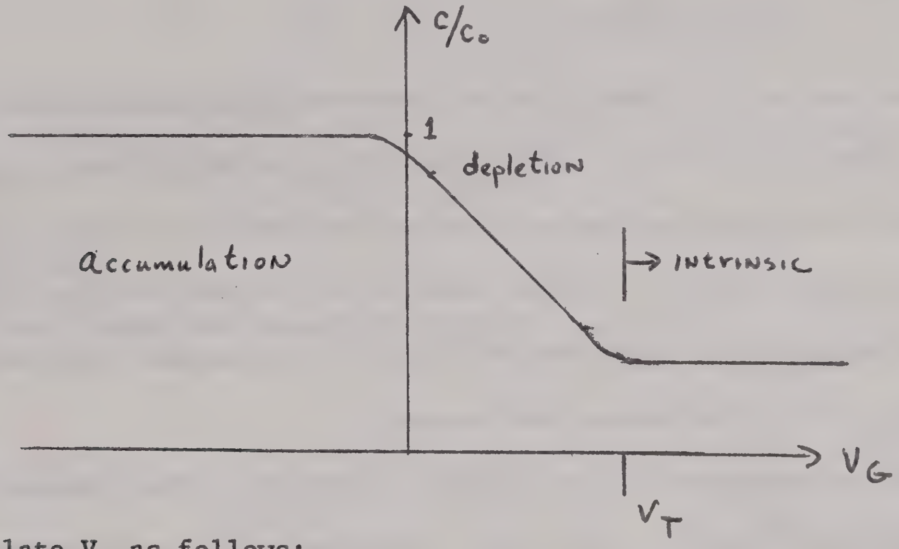
$$\frac{C_{ox}}{C} = 1 - \frac{\varepsilon_{ox}}{x_{ox}}\frac{1}{\varepsilon_{Si}}\left(\frac{2\varepsilon_{Si}}{qN_A}\right)^{\frac{1}{2}}\frac{qN_A}{2}\frac{x_{ox}}{\varepsilon_{ox}}\left(\frac{2\varepsilon_{Si}}{qN_a}\right)^{\frac{1}{2}}$$

$$+ \left[1 + 2\frac{\varepsilon_{ox}^2}{x_{ox}^2}\frac{1}{qN_A\varepsilon_{Si}} V_G\right]^{\frac{1}{2}}$$

Hence

$$(9) \qquad \frac{C}{C_{ox}} = \frac{1}{\left[1 + \frac{2\varepsilon_{ox}^2}{qN_A\varepsilon_{Si}x_{ox}^2} V_G\right]^{\frac{1}{2}}}$$

This equation is valid in the depletion region. If $V_G < 0$ we have an accumulation condition and the silicon acts as a lossy conductor. Hence $\frac{C}{C_{ox}} \simeq 1$ if $V_G < 0$. If $V_G > 0$ the depletion region forms and is voltage dependent. The effective capacitance C decreases. But x_0 reaches a maximum value. This occurs when $V_G = V_T$. If $V_G > V_T$, i.e. in the inversion region, the capacitance will again be voltage independent.



We may calculate V_T as follows:

(1a) $$V_G = -\frac{Q_S}{C_{ox}} + V_S$$

But at the onset of inversion

$$V_S = V_{Si} = 2 + 2\,\frac{kt}{q}\,\ln\frac{N_A}{n_i}$$

and $$Q_S = -\,q\,N_A\,x_{DMAX} = -\,9N_A\left[\frac{2\varepsilon_{Si}}{qN_A}\,V_{Si}\right]^{\frac{1}{2}}$$

Hence

(10) $$V_T = qN_A\,\frac{x_{ox}}{\varepsilon_{ox}}\left[\frac{2\varepsilon_{Si}}{qN_A}\;V_{Si}\right]^{\frac{1}{2}} + V_{Si}$$

We can therefore control this voltage via the oxide thickness and the background doping. There is however an additional constraint which we must impose on the capacitance measurement in the depletion mode. This comes about from the very fact that we do have a depletion region which in turn implies a thermal generation rate. This rate is (See page 75)

$$G = \frac{n_i}{2T}$$

it contributes in time t

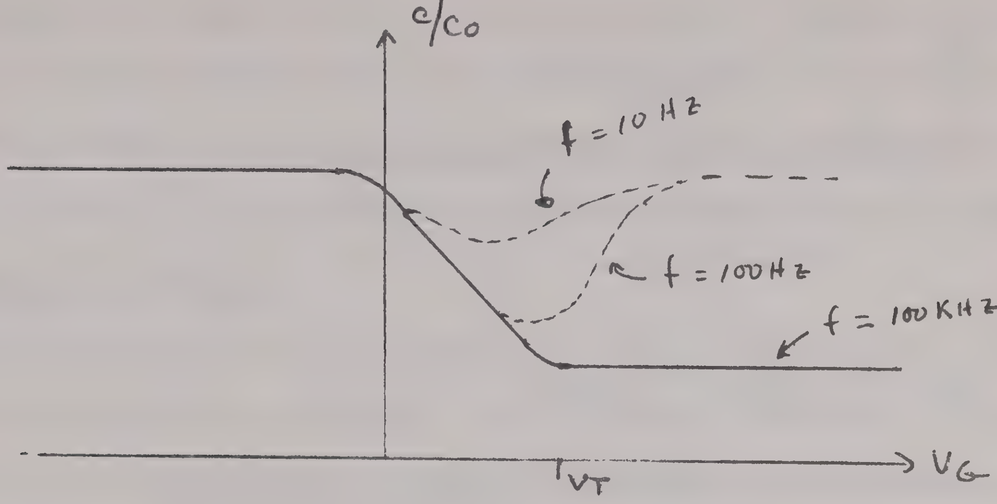
$$Q = q \frac{n_i x_D}{2T} At$$

electron-hole pairs. Here A is the cross-sectional area of the device. In the steady state, i.e. if V_G = constant we would therefore have to admit to the fact that the depletion region contains some mobile carriers, in particular, since the field for $V_G > 0$ would tend to pull holes into the bulk, i.e. away from the surface the electrons would tend to accumulate near the interface. This idea has the following implication:

Suppose we change $V_G \to V_G + \Delta V_G$ in time Δt. This changes the charge on the metal gate from $Q_G \to Q_G + \Delta Q_G$. The charge ΔQ_G is compensated in the semiconductor as follows:

(a) if $\Delta Q_G >> A \frac{qn_i}{2T} x_D \Delta t$ the depletion region will increase from $x_D \to x_D + \Delta x_D$ and will supply $\Delta Q_G = -qN_A A \Delta x_D$ charge to achieve over-all charge neutrality. This is the normal depletion layer behavior. Our earlier capacitance curve would be correct. Note that it requires Δt to be small.

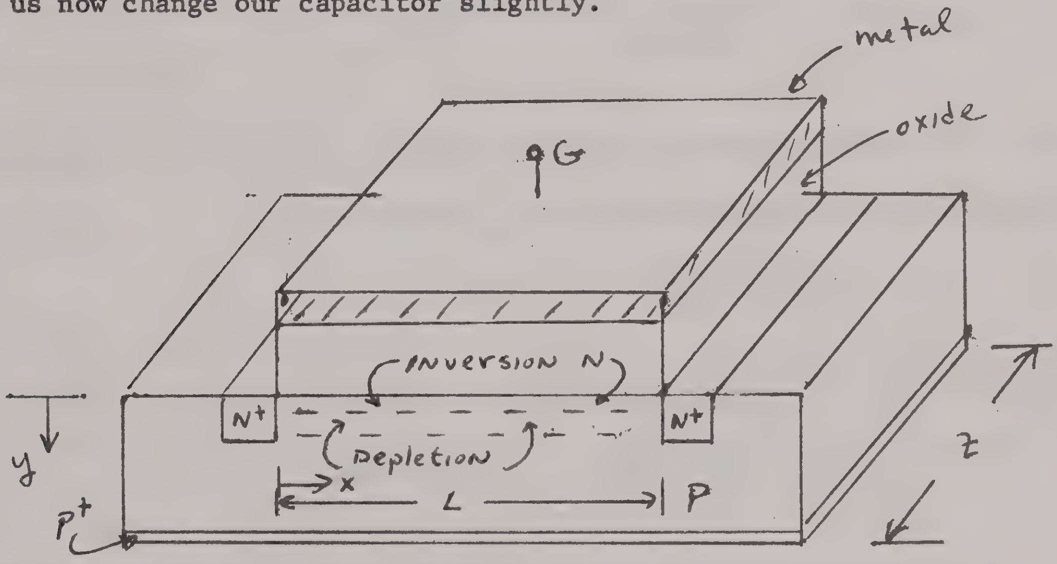
(b) If $\Delta Q_G << A \frac{qn_i}{2T} x_D \Delta t$ the depletion region will not increase in size. Electrons will be swept to the interface until ΔQ_G is compensated by electrons in an inversion mode. This, however, states that for slow changes the depletion region does not enter into the capacitance. The oxide capacitance only should be measured if Δt is long. Our results are frequency dependent.



Typical results are shown in the above figure.

CHANNEL CONDUCTANCE

Let us now change our capacitor slightly.



Let us supply a voltage V_G between gate and ground. We then supply two N^+-contacts which make ohmic contact to the inversion layer if it exists. The conductivity in the y - direction is

$$\sigma = q\mu_n n(y) + q\mu_p p(y)$$

But there are no holes to the surface if $V_G > 0$. Hence

$$\sigma = q\mu_n n(y)$$

n(y) is large near the surface and decays to n_i at y_i if $y > y_i$ we are in the depletion region. The average channel conductivity is therefore

$$(11) \qquad \langle\sigma\rangle = \frac{1}{y_i}\int_0^{y_i} \sigma(y)\, dy$$

the resistance between the two N^+-contacts is then

$$R = \frac{1}{\langle\sigma\rangle}\ \frac{L}{Zy_i}$$

or

$$(12) \qquad G = \frac{1}{R} = \frac{Z}{L}\int_0^{y_i} \sigma(y)dy$$

if we assume that the mobility is constant we can write

$$Q_n = -q\int_0^{y_i} n\, dy \qquad \text{coulombs/unit area}$$

$$(13) \qquad G = -\frac{Z}{L}\mu_n Q_n$$

But if the total charge in the semiconductor is Q_S and the depletion layer charge in the inversion region is Q_B we can write

$$Q_n = Q_S - Q_B$$

But according to Eq. (10) page 97

$$V_T = -\frac{Q_B}{C_{ox}} + V_{Si} \qquad \text{(inversion)}$$

$$-Q_B = C_{ox}\ (V_T - V_{Si})$$

and $$Q_S = C_{ox}\ (V_S - V_G)$$

Hence, $Q_N = C_{ox}\ (V_S - V_G + V_T - V_{Si})$

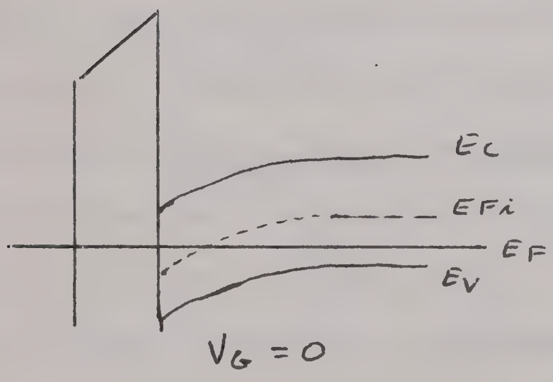
$$(14) \qquad G = +\frac{Z}{L}\mu_n\ (V_G - V_S - V_T + V_{Si})\ C_{ox}$$

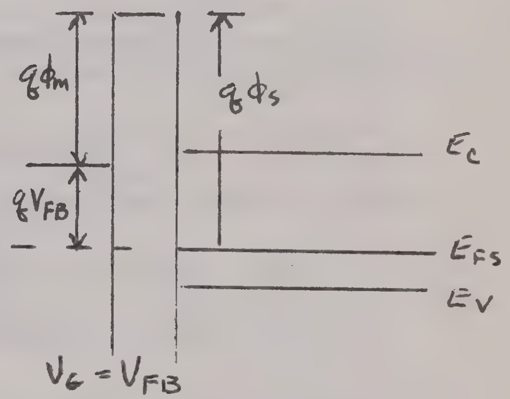
$$\simeq \frac{Z}{L}\mu_n\ (V_G - V_T)\ C_{ox} \quad \text{if } V_S \simeq V_{Si}$$

This suggests that the conductance is zero until $V_G = V_T$ or until the inversion regime begins. We might therefore call V_T the turn-on voltage and agree that it is of technical importance for a device. The entire mos structure acts very much like the JFET: it is a voltage controlled resistor. Note, however, that the N-P junction is field-induced rather than mechanical as for the JFET.

WORK FUNCTION DIFFERENCES

The electron energies at the fermi level of the metal and semiconductor in the mos structure are different. This difference is normally expressed in terms of the work functions of the two materials: The energy which is needed to move an electron from the fermi level to vacuum. Contact between metal-oxide-semiconductor leads to a charge exchange between the metal and the semiconductor. A field will appear across the oxide. The resulting band structure is as shown





If V_G is adjusted to V_{FB} the energy diagram will be as shown on the right. We call V'_{FB} the flat band voltage. It is equal to

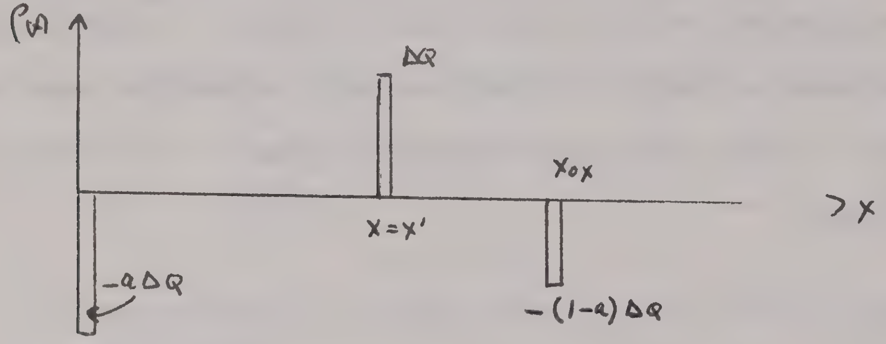
$$V'_{FB} = \phi_M - \phi_S \equiv \phi_{MS}$$

Note that ϕ_M and ϕ_S are modified work functions: The energy required to move an electron from the fermi level to the conduction band of the oxide. Both quantities are material parameters. Hence V_{FB} is a constant for say an aluminum-oxide-silicon structure.

INSULATOR CHARGES

Our analysis of the mos structure was based on the assumption that we have a perfect dielectric. Normally we produce the dielectric by thermal oxidation of silicon. This implies that it may contain some silicon ions S_i^+, some oxygen ions 0^- and a variety of inpurity ions N_a^+, K^+ etc.. These result in a charged oxide. The question which concerns us here involves their effect on the mos structure.

Consider a charge ΔQ which is somewhere in the oxide. It induces the charge $-a\Delta Q$ in the metal and $-(1-a)\Delta Q$ in the semiconductor.



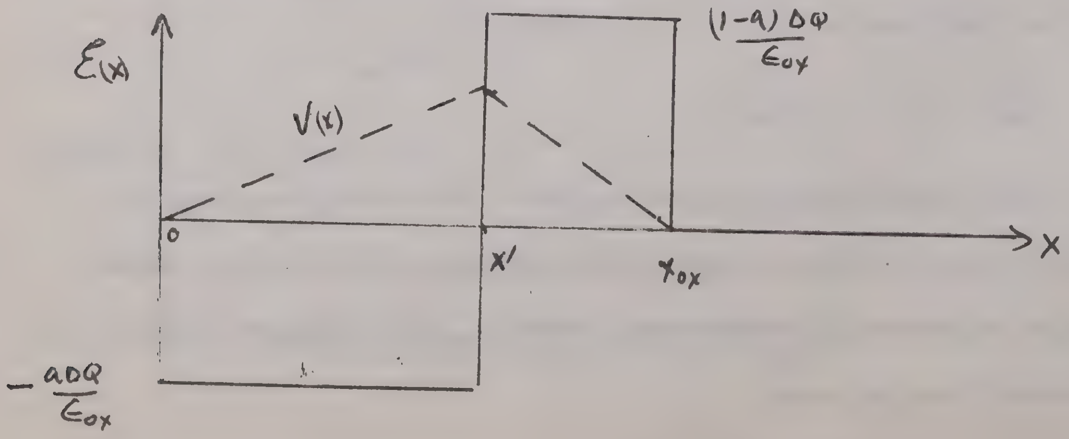
Then, assuming delta functions

$$\varepsilon_{ox}\frac{dE}{dx} = -a\Delta Q\delta(x) + \Delta Q\delta(x-x') - (1-a)\ \Delta Q\delta(x-x_{ox})$$

$$E = -\frac{a\Delta Q}{\varepsilon_{ox}} \quad 0 < x < x'$$

$$= \frac{(1-a)\Delta Q}{\varepsilon_{ox}} \quad x' < x < x_{ox}$$

the field is therefore as shown:



The potential is related to the field by

$$\frac{dV}{dx} = - E$$

Hence $$\int_{V(0)}^{V(x_{ox})} dV = \frac{a\Delta Q}{\varepsilon_{ox}} x' - \frac{1-a}{\varepsilon_{ox}} \Delta Q \, (x_{ox}-x') = 0$$

But this implies that

$$ax' - (1-a)\,(x_{ox}-x') = 0$$

or $$a = \frac{x_{ox}-x'}{x_{ox}}$$

The two induced charges are therefore

$$Q_M = -(1 - \frac{x'}{x_{ox}})\Delta Q \qquad \text{(Metal)}$$

and $$Q_S = - \frac{x'}{x_{ox}} \Delta Q \qquad \text{(Semiconductor)}$$

We now claim that the only effect of ΔQ is that of shifting the flat band voltage. In other words: the flatband condition requires that the semiconductor be charge neutral. Thus, if we apply a voltage V_{ext} such that this external voltage causes $\hat{Q}_S = \frac{x'}{x_{ox}} \Delta Q$ to be induced in the semiconductor then the total Q_S will be zero. But this is the flatband condition precisely. The shift in the flatband voltage may be calculated by considering the charge density:

$$\hat{\rho}(x) = - \frac{x'}{x_{ox}} \Delta Q \delta(x) + \frac{x'}{x_{ox}} \Delta Q \delta(x-x_{ox})$$

This yields

$$E = - \frac{x'}{\varepsilon_{ox} x_{ox}} \Delta Q$$

$$V(x_{ox})-V(0) = \frac{x'}{\varepsilon_{ox} x_{ox}} \Delta Q \, x_{ox}$$

But $V_G = V(0) - V(x_{ox})$

Hence

$$V_G = -\frac{x' \Delta Q}{\varepsilon_{ox}} = -\frac{x'}{x_{ox}} \frac{\Delta Q}{C_{ox}}$$

Since this is the shift in the flatband voltage we denote it by V_{FB}. Normally the oxide does not only contain ΔQ but the charge density $\rho(x)$ in the oxide. The total shift in the flat band voltage become

$$V_{FB}'' = -\frac{1}{C_{ox}} \int_0^{x_{ox}} \frac{x\rho(x)}{x_{ox}} dx$$

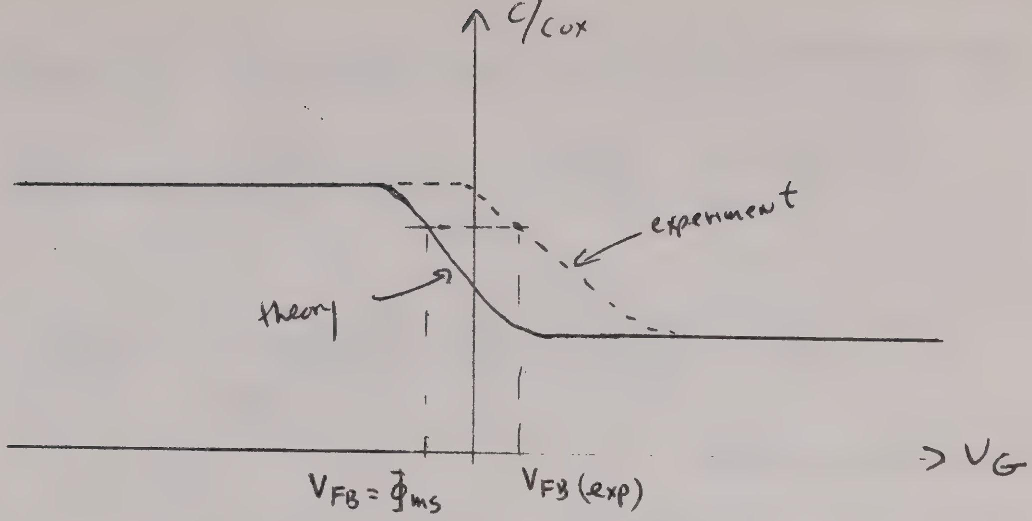
The contact between the silicon and silicon dioxide represents a discontinuity in the silicon energy diagram. Dangling surface bonds contribute a charge Q_{SS}. Its contribution to the flat-band voltage shift is

$$V_{FB}''' = -\frac{Q_{SS}}{C_{ox}}$$

We add also the work function difference to obtain the total flatband voltage as

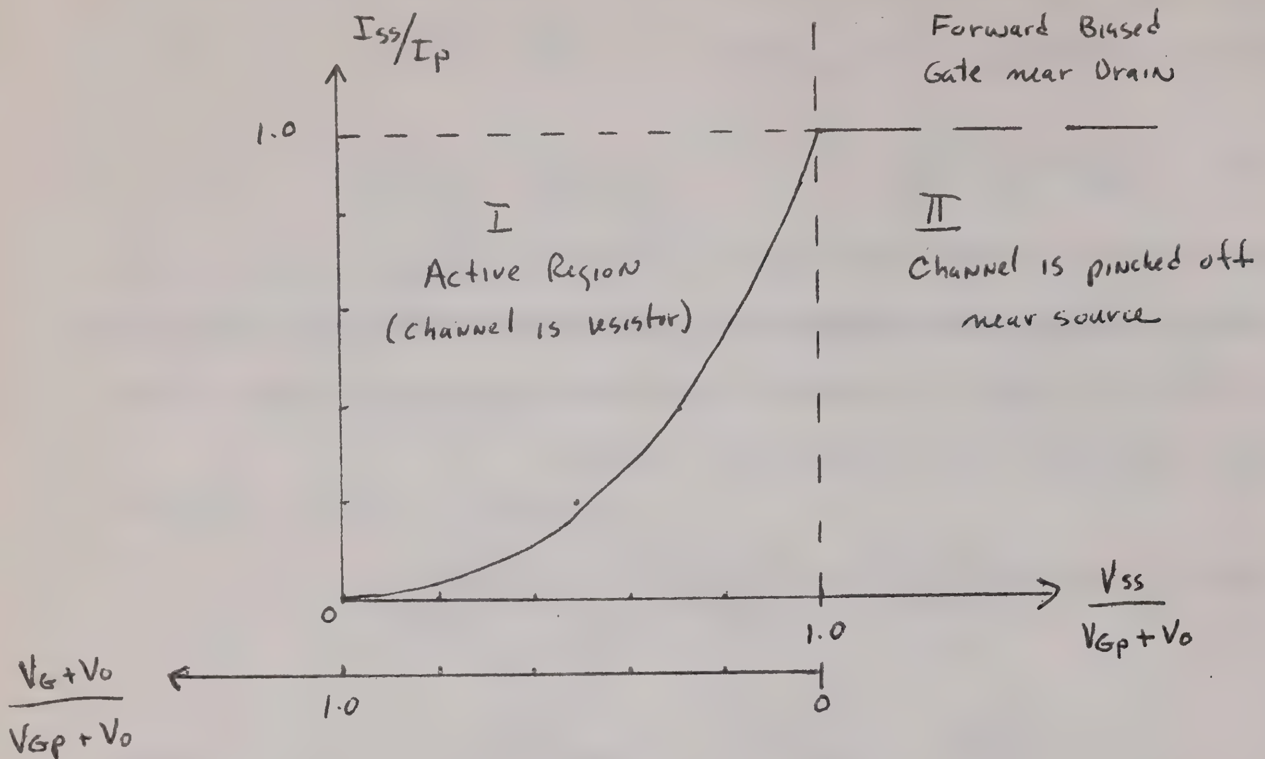
$$V_{FB} = \Phi_{MS} - \frac{Q_{SS}}{C_{ox}} - \frac{1}{C_{ox}} \int_0^{x_{ox}} \frac{x\rho(x)}{x_{ox}} dx$$

The effect of the flat band voltage is essentially that of shifting the high frequency C-V characteristic along the V_G axis. Thus, if we assume a perfect dielectric and no surface charge we calculate an ideal C-V curve. A measured C-V curve will then appear as shown.



It is of course obvious that the shift in the C-V characteristic will also cause the turn-on voltage V_T to be shifted. Matter of fact, if the oxide charges are sodium ions the application of a field will cause ion drift which changes $\rho(x)$, which changes V_{FB} and therefore V_T. The turn on point would become time dependent. The resulting device could not be biased in a stable fashion. It is exactly this problem which had to be overcome in order to make mos devices practical.

The saturation characteristic is shown in the next figure



The region of validity for our analysis is region I: The channel is a controlled resistor. In this region Eq. (9b) applies. Since

$$\frac{V_G+V_0}{V_{GP}+V_0} = 1 - \frac{V_{SS}}{V_{GP}+V_0}$$

and

$$3I_P = \frac{V_{GP}+V_0}{R_0}$$

We may write (9b) as:

$$(10)\quad \frac{1}{3}\,\frac{I_S}{I_P} = 1 - \frac{V_{SS}-V_S}{V_{GP}+V_0} - \frac{2}{3}\left[1 - \frac{V_{SS}-V_S}{V_{GP}+V_0}\right]^{3/2} - \left[1 - \frac{V_{SS}}{V_{GP}+V_0} - \frac{2}{3}\left[1- \frac{V_{SS}}{V_{GP}+V_0}\right]^{3/2}\right.$$

or

$$(11)\quad \frac{I_S}{I_P} = 2\left[1 - \frac{V_{SS}}{V_{GP}+V_0}\right]^{3/2} -2\left[1- \frac{V_{SS}-V_S}{V_{GP}+V_0}\right]^{3/2} +3\,\frac{V_S}{V_{GP}+V_0}$$

This equation becomes Eq. (9e) if we let $V_S \to V_{SS}$. It describes the active where $V_S + V_G \le V_{GP}$ and $V_G + V_0 \ge 0$

Suppose now that we set V_S to V_{SS} where

$$(10) \qquad V_{SS} = V_{GP} - V_G$$

The channel will just pinch off at $x = 0$. Further increases in V_S, i.e. above V_{SS}, would no longer cause increases in I_S since the source contact itself would be in a dedletion region. We speak of a saturated condition and call (10) the saturation condition. At saturation (9b) becomes:

$$(9c) \qquad 1 - \frac{2}{3} - \left[\frac{V_G + V_0}{V_{GP}+V_0} - \frac{2}{3} \left[\frac{V_G+V_0}{V_{GP}+V_0} \right]^{3/2} \right] = \frac{I_{SS}R_0}{V_{GP}+V_0}$$

or

$$1 - 3\,\frac{V_G+V_0}{V_{GP}+V_0} + 2 \left(\frac{V_G+V_0}{V_{GP}+V_0} \right)^{3/2} = I_{SS}\,\frac{3R_0}{V_{GP}+V_0}$$

Note that $\quad I_P = \dfrac{V_{GP}+V_0}{3R_0}$

is a device constant. The saturation condition becomes

$$(9d) \qquad I_{SS} = I_P \left\{ 1-3\,\frac{V_G+V_0}{V_{GP}+V_0} + 2 \left(\frac{V_G+V_0}{V_{GP}+V_0} \right)^{3/2} \right\} V_G \ge -V_0$$

or, since (10) may be written as

$$\frac{V_{SS}}{V_{GP}+V_0} = 1 - \frac{V_G+V_0}{V_{GP}+V_0} \qquad \frac{V_G+V_0}{V_{GP}+V_0} = 1 - \frac{V_{SS}}{V_{GP}+V_0}$$

$$(9e) \qquad \frac{I_{SS}}{I_P} = 2 \left(\frac{V_G+V_0}{V_{GP}+V_0} \right)^{3/2} - 3\,\frac{V_G+V_0}{V_{GP}+V_0} \rightarrow + 1$$

$$(9e) \qquad \frac{I_{SS}}{I_P} = 2 \left[1 - \frac{V_{SS}}{V_{GP}+V_0} \right]^{3/2} + 3\,\frac{V_{SS}}{V_{GP}+V_0} - 2$$

or, in the unit

$$\eta(x)\,\frac{dV}{dx} = \frac{I_S \rho_N}{2T}$$

or

$$\text{(7)}\qquad \frac{\eta(x)}{a}\,\frac{dV}{dx} = \frac{I_S \rho_N}{2aT} = \frac{I_s}{W}\,R_0$$

Here R_0 is the channel resistance. It is a device constant.

$$R_0 = \rho_N \frac{W}{2aT}$$

combining (7) and (6) yields

$$\text{(8)}\qquad \left\{1-\left[\frac{V_S + V_G + V_0 - V(x)}{V_{GP} + V_0}\right]^{\frac{1}{2}}\right\}\frac{dV}{dx} = \frac{I_s}{W}\,R_0$$

where

$$V(0) = 0$$

$$V(W) = V_s - V_D$$

Equation (8) is the basic equation for the junction FET. It's integration is straight forward and yields:

$$\text{(9)}\qquad V_S - V_D + \frac{2}{3}\,\frac{1}{\sqrt{V_{GP}+V_0}}\left[(V_D + V_G + V_0)^{3/2} - (V_S+V_G+V_0)^{3/2}\right] = I_S R_0$$

or

$$\text{(9a)}\qquad \frac{V_S - V_D}{V_{GP}+V_0} + \frac{2}{3}\left[\left(\frac{V_D+V_G+V_0}{V_{GP}+V_0}\right)^{3/2} - \left(\frac{V_S+V_G+V_0}{V_{GP}+V_0}\right)^{3/2}\right] = \frac{I_S R_0}{V_{GP}+V_0}$$

Examination of the JFET diagram indicates that we may set V_D to zero without loss of generality. This yields

$$\text{(9b)}\qquad \frac{V_S+V_G+V_0}{V_{GP}+V_0} - \frac{2}{3}\left[\frac{V_S+V_G+V_0}{V_{GP}+V_0}\right]^{3/2} - \left[\frac{V_G+V_0}{V_{GP}+V_0} - \frac{2}{3}\left[\frac{V_G+V_0}{V_{GP}+V_0}\right]^{3/2}\right] = \frac{I_S R_0}{V_{GP}+V_0}$$

$$(3) \qquad V_{NP}(x) = -V(x) + V_S + V_G + V_0$$

note that

$$(3a) \qquad V_{NP}(0) = V_S + V_G + V_0 \qquad (3b) \quad V_{NP}(W) = V_D + V_G + V_0$$

in order for the current I_S to flow $V_S > V_D$. The maximum reverse bias occurs therefore at $x = 0$. But then x_N is a maximum near $x = 0$. If we want to prevent pinch-off for any x we insist on

$$(4) \qquad V_{NP}(0) \leq V_{GP} + V_0 \qquad V_S + V_G \leq V_{GP}$$

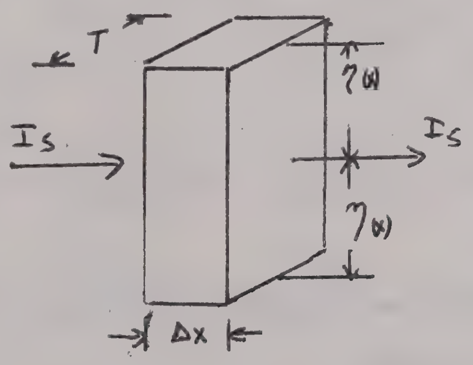
on other hand, if we do not want to forward bias the channel any place

$$(5) \qquad V_{NP}(W) \geq 0 \qquad V_D + V_G + V_0 \geq 0$$

These two inequalities estabilish the limits of the validity range of our analysis. If they are satisfied the following analysis holds

$$(6) \qquad \eta(x) = a - x_N(x) = a\left(1 - \left[\frac{V_S + V_G + V_0 - V(x)}{V_{GP} + V_0}\right]^{1/2}\right)$$

The current I_S causes the voltage drop $V(x)$. Consider a slice through the channel

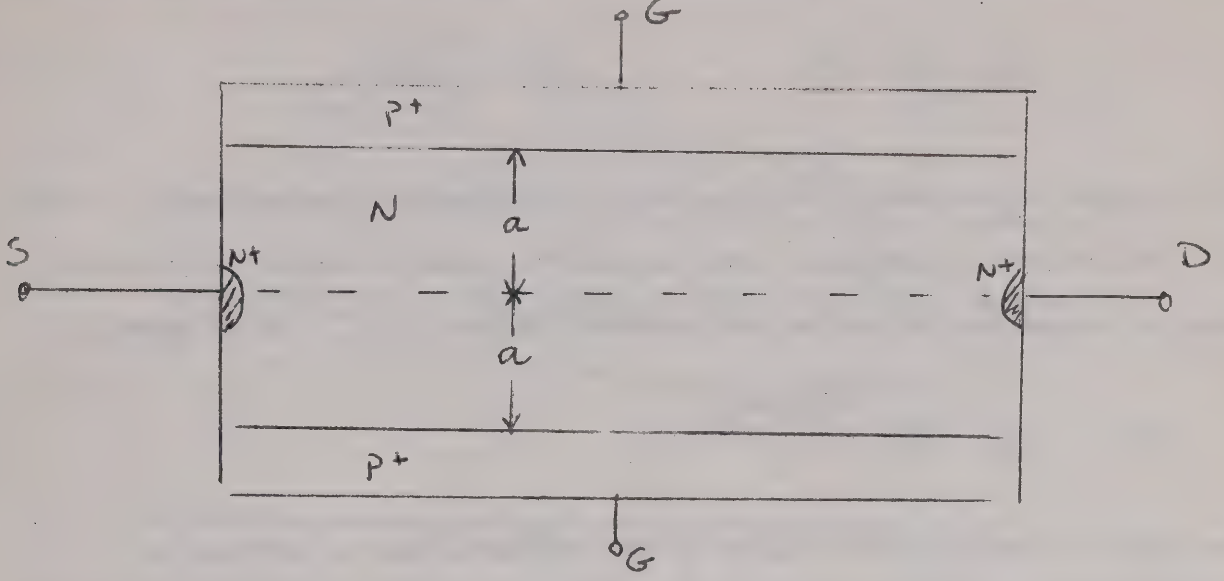


the voltage drop is

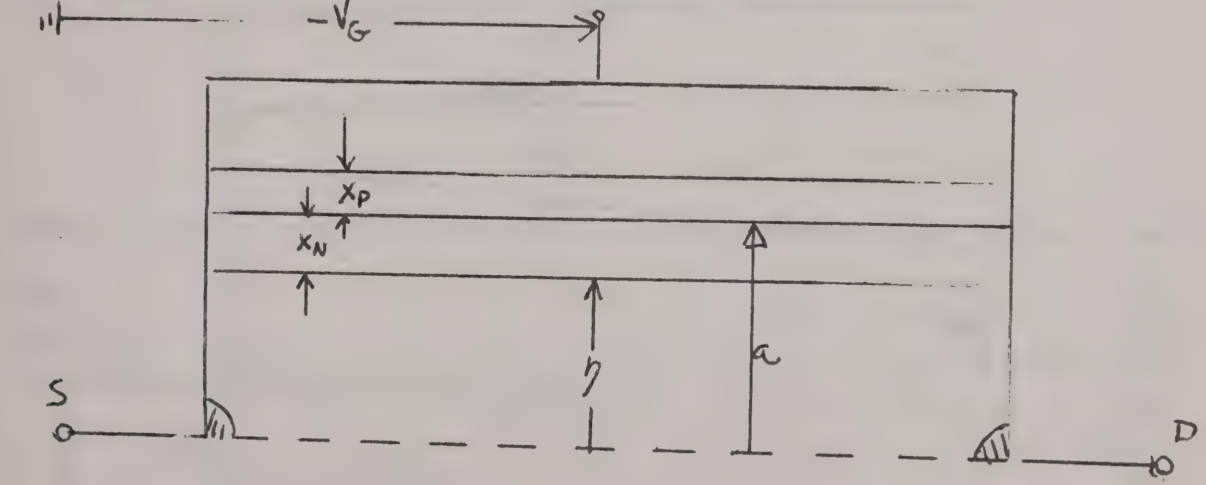
$$\Delta V = I_S \, \rho_\eta \frac{\Delta x}{2\,\eta(x)\,T}$$

THE JUNCTION FET

Suppose we combine the idea of a resistor and a depletion region.



The P^+-region is a heavily doped contact region. The N^+ contacts are ohmic contacts between S and D. Hence the network element between S and D is a resistor. If we reverse bias the P^+-N diodes a depletion region forms which will increase the resistance S-D. Consider grounding S and D. Then we have the following



if the doping profile is abrupt the value of x_N will be

$$x_N = \left(\frac{2\varepsilon V_{NP}}{q} \; \frac{N_A}{N_D(N_A+N_D)} \right)^{\frac{1}{2}}$$

$$V_{NP} = V_0 + V_G \qquad\qquad V_0 = \frac{kT}{q} \,\text{lu}\, \frac{N_A N_D}{n_i^2}$$

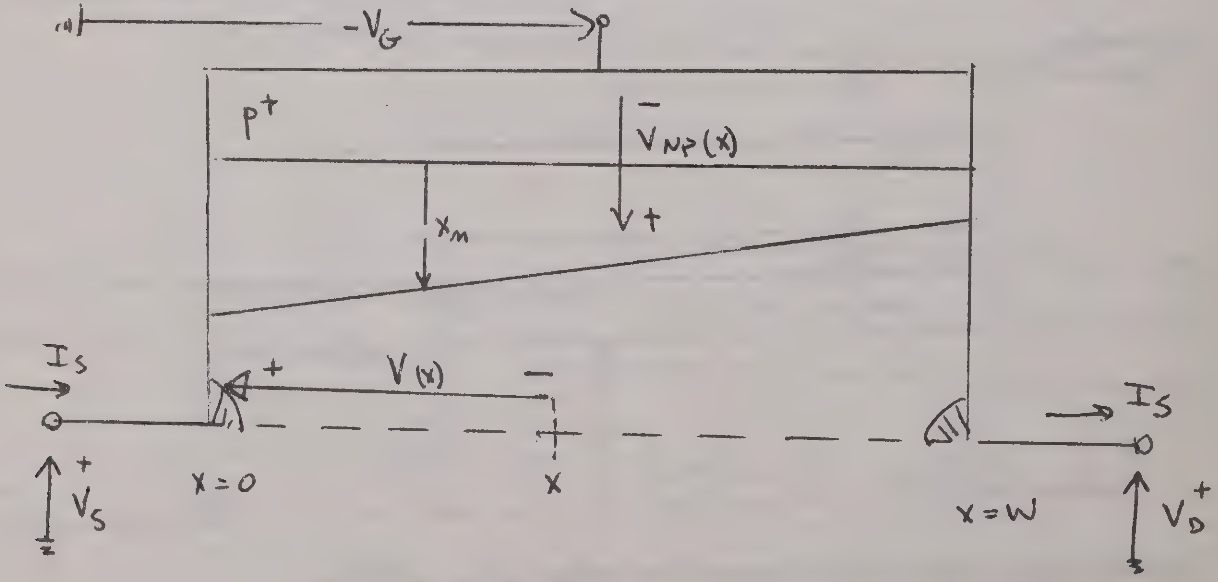
If we adjust V_G so that $a = x_N$ the entire N-region will be depleted. We call this the pinch-off condition. The value of V_G which is needed to pinch the bulk or channel off is called the pinch-off potential V_{GP}. It is given by

$$(1) \qquad V_{GP} = \frac{qa^2}{2\varepsilon} \; \frac{N_D(N_A+N_D)}{N_A} - V_0$$

The width of the depletion region in the N-material may be written as

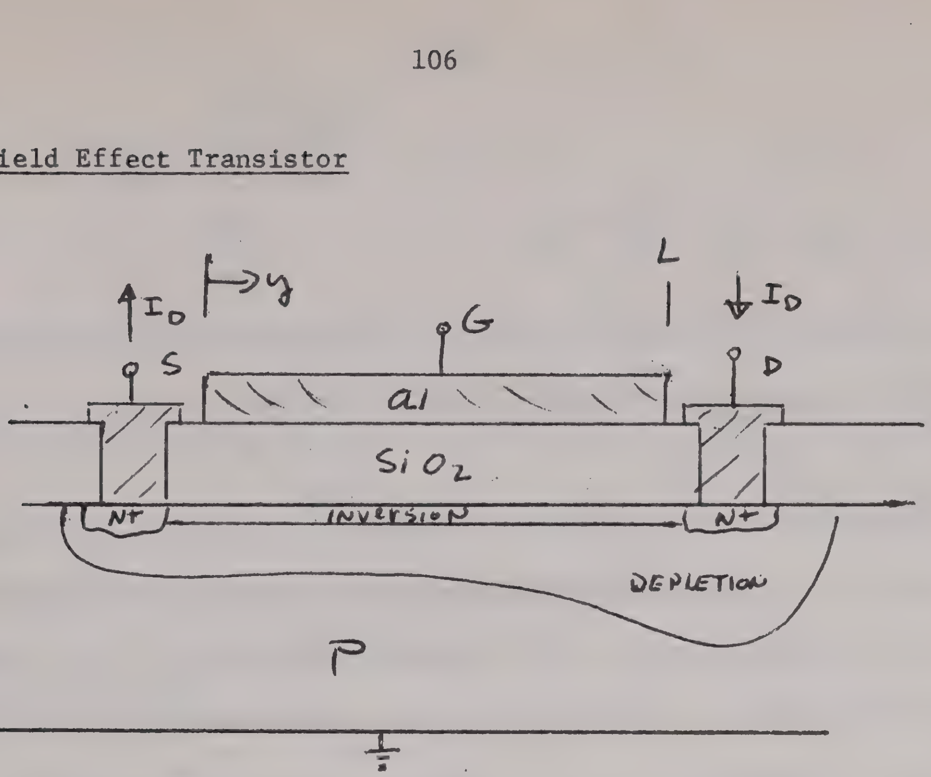
$$(2) \qquad x_N = a \sqrt{\frac{V_G+V_0}{V_{GP}+V_0}} = a \sqrt{\frac{V_{NP}}{V_{GP}+V_0}}$$

Suppose now that we remove the ground from S and D. A current will flow and the following diagram will apply:

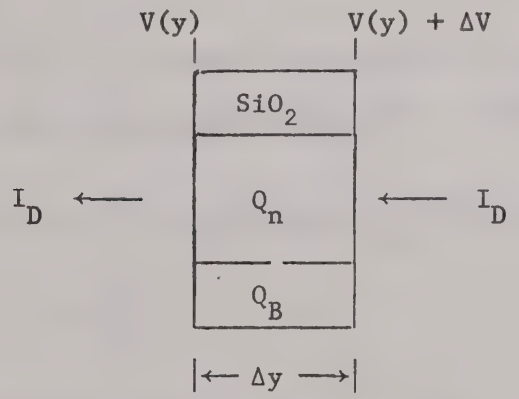


The current I_S causes the channel voltage drop $V(x)$. Note that we are neglecting the reverse saturation current of the diode. The total potential across the depletion region becomes

Surface Field Effect Transistor



In the above sketch we describe a MOSFET in the active region. If we consider a differential cross-section Δy we find



Hence Q_n is the inversion charge and Q_B the depletion layer charge. The voltage drop

$$(1) \quad \Delta V = -I_D \, dR = -I_D \frac{\Delta y}{Z\mu_n Q_n(y)}$$

or $$\frac{dV}{dy} = -\frac{I_D}{Z\mu_n Q_n(y)} \ .$$

The total semiconductor charge is

$$(2) \quad Q_s = Q_n + Q_B$$

and

$$(3) \quad V_G - V_{FB} = -\frac{Q_{si}}{C_{ox}} + V_s \; .$$

The last equation is Eq. (1a) on page 96. It has been modified to include the flatband voltage. Combining (2) and (3) yields:

$$(4) \quad Q_n(y) = -[V_G - V_{FB} - V_s]C_{ox} - Q_B(y) \; .$$

The voltage V(y) is the voltage measured from a point in the channel inversion layer to the grounded source or the grounded P-substrate. It is therefore also the voltage which is measured from the semiconductor surface to the neutral P-bulk. We write

$$(5) \quad V_s(y) = V(y) + V_{si}$$

where V(y) is the voltage across the depletion region. V_{si} is of course the potential at which inversion starts. The depletion layer charge becomes

$$(6) \quad Q_B(y) = -qN_A x_D(y) = -\left[2\varepsilon_{si} qN_A\left(V(y)+V_{si}\right)\right]^{1/2}$$

and

$$(4a) \quad Q_n(y) = -[V_G - V_{FB} - V(y) - V_{si}]C_{ox} + \left|2\varepsilon_{si} qN_A\left(V(y)+V_{si}\right)\right|^{1/2} \; .$$

The differential equation for the device becomes

$$(7) \quad I_D = Z\mu_n C_{ox}\left[V_G - V_{FB} - V_{si} - V(y) - \frac{1}{C_{ox}}\left[2\varepsilon_{si} qN_A\left(V_{si}+V(y)\right)\right]\right]^{1/2} \frac{dV}{dy} \; .$$

The integration is readily performed. V(0) = 0 and V(L) = V_D yields

$$(8) \quad I_D = \frac{Z}{L}\mu_n C_{ox}\left[\left(V_G - V_{FB} - V_{si} - \frac{V_D}{2}\right)V_D - \frac{2}{3}\frac{\sqrt{2\varepsilon_{si} qN_A}}{C_{ox}}\left[(V_D+V_{si})^{3/2} - V_{si}^{3/2}\right]\right]$$

This is the characteristic of the MOSFET. It may be expressed slightly differently by noting that

$$x^{\circ}_{DMAX} = \sqrt{\frac{2\varepsilon_{si}V_{si}}{qN_A}}$$

$$C_{ox} = \frac{\varepsilon_{ox}}{x_{ox}} \ .$$

This yields

$$(8a) \quad I_D = \frac{Z}{L}\mu_n C_{ox}\left[\left(V_G - V_{FB} - V_{si} - \frac{V_D}{2}\right)V_D - \frac{4}{3}\frac{\varepsilon_{si}x_{ox}}{\varepsilon_{ox}x_{DMAX}}\sqrt{V_{si}}\left[(V_D + V_{si})^{3/2} - V_{si}^{3/2}\right]\right]$$

This equation is of course valid only if the device is not saturated: the inversion charge exists along the entire channel.

Approximate response:

If $V_D < < V_{si}$

$$(8c) \quad I_D \approx \frac{Z}{L}\mu_n C_{ox}(V_G - V_{FB} - V_{si})V_D$$

which we may write as

$$I_D = \frac{Z}{L}\mu_n C_{ox}(V_G - V_T')V_D \qquad V_T' = V_{FB} + V_{si} \ .$$

The transconductance of the device is defined by

$$(9) \quad g_m = \frac{\partial I_D}{\partial V_G}$$

$$= \frac{Z}{L}\mu_n C_{ox} V_D$$

and the channel conductance by

$$(10) \quad g = \frac{\partial I}{\partial V_D}$$

$$= \frac{Z}{L}\mu_n C_{ox}(V_G - V_T)$$

where $V_G > V_T$ and $V_D < V_{DSAT}$.

The saturation condition is specified by

$$Q_N(L) = 0.$$

The inversion region disappears near the drain. Since

$$(4b)\quad Q_N(L) = -\,[V_G - V_{FB} - V(L) - V_{si}]C_{ox} + \left[2\varepsilon_{si}qN_A\left(V(L)+V_{si}\right)\right]^{1/2}$$

we obtain at saturation

$$0 = -\,[V_G - V_{FB} - V_{Ds} - V_{si}]C_{ox} + [2\varepsilon_{si}qN_A(V_{si}+V_{Ds})]^{1/2}\ .$$

If this equation is solved for V_{Ds} Eq. (11) results

$$(11)\quad V_{Ds} = V_G - V_{FB} - V_{si} + \frac{\varepsilon_{si}qN_A}{C_{ox}^2}\left[1 - \left[1 + \frac{2C_{ox}^2(V_G - V_{FB})}{\varepsilon_{si}qN_A}\right]^{1/2}\right]\ .$$

This equation, subject to the condition $x_{DMAX} >> x_{ox}$, approximates to

$$(12)\quad V_{Ds} \simeq V_G - V_{FB} - V_{si}\ .$$

Transport theory:

The devices which we have considered thus far involve the concept of a depletion region: no mobile charges, and the concept of a resistor: no excess mobile charges. In order to treat more general devices we must consider the behavior of excess densities in the semiconductor.

We know:

$$(1)\quad J_N = q\ {}_n nE + qDn\,\frac{dn}{dx}$$

$$(2)\quad J_P = q\ {}_p pE - qD_p\,\frac{dn}{dx}$$

$$(3)\quad J = J_N + J_P$$

$$(4)\quad \frac{dE}{dx} = \frac{q}{\varepsilon}\,(p - n + N_D^{\ +} - N_A^{\ -})\ .$$

The equation which we have not used yet is

$$(5) \quad \overline{\nabla} \cdot \overline{J} + \frac{\partial \hat{\rho}}{\partial t} = 0$$

this is simply a statement of charge conservation. Note that $\hat{\rho}$ refers to the mobile charge: $q(p-n)$. We might write

$$(5a) \quad \overline{\nabla} \cdot (\overline{J}_N + \overline{J}_P) + q \frac{\partial}{\partial t} (p-n) = 0$$

$$\left(\overline{\nabla} \cdot \overline{J}_N - q \frac{\partial n}{\partial t}\right) + \left(\overline{\nabla} \cdot \overline{J}_P + q \frac{\partial p}{\partial t}\right) = 0 \quad .$$

The first term refers only the the electrons, the second only to the holes. These two terms are also conservation statements. Note, however, that we can generate an electron and a hole in particular if we say that we have a pair generation rate qG and a pair recombination rate qR then the corresponding hole charge will be $q(G-R)$. Hence

$$(6) \quad \overline{\nabla} \cdot \overline{J}_P + q \frac{\partial p}{\partial t} = q(G-R)$$

and

$$(7) \quad \overline{\nabla} \cdot \overline{J}_N - q \frac{\partial n}{\partial t} = -q(G-R) \; .$$

Recall that G may contain an external generation rate as well as the thermal generation rate. If the external rate is zero we know that

$$(8) \quad G - R = -\frac{p - p_0}{\tau} = -\frac{n - n_0}{\tau}$$

and

$$(6a) \quad \frac{\partial p}{\partial t} = -\frac{p - p_0}{\tau} - \frac{1}{q} \frac{\partial J_P}{\partial x}$$

$$(7a) \quad \frac{\partial n}{\partial t} = -\frac{n - n_0}{\tau} + \frac{1}{q} \frac{\partial J_N}{\partial x} \quad .$$

Note that we have implied that

$$n - n_0 = p - p_0 \qquad \text{or} \qquad \Delta n = \Delta p$$

$$n = p - p_0 + n_0 \; .$$

If we substitute (2) into (6a) we get

$$\frac{\partial p}{\partial t} = - \frac{p - p_0}{\tau} - \frac{1}{q} \frac{\partial J_p}{\partial x}$$

$$(8) \qquad \frac{\partial p}{\partial t} = - \frac{p - p_0}{\tau} - \mu_p \mathcal{E} \frac{\partial p}{\partial x} - \mu_p p \frac{\partial \mathcal{E}}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2}$$

$$\text{but} \qquad (4) \qquad \frac{\partial \mathcal{E}}{\partial x} \approx \frac{q}{\varepsilon}\left(p - (p - p_0 + n_0) + N_D^{\;+} - N_A^{\;-}\right)$$

$$\approx \frac{q}{\varepsilon} (p_0 - n_0 + N_D^{\;+} - N_A^{\;-}) \; .$$

This equation suggests that if the semiconductor was charge neutral at equilibrium: the doping porfile did not change too rapidly, the quantity $\frac{\partial \mathcal{E}}{\partial x}$ in non-equilibrium will be small. But Eq. (8) contains the term $p \frac{\partial \mathcal{E}}{\partial x}$. We may then argue that this term is small if we select p as the minority carrier. If p is the majority carrier than $p \frac{\partial \mathcal{E}}{\partial x}$ is not small. However, in that case $n \frac{\partial \mathcal{E}}{\partial x}$ will be small. This suggests that we can approximate the minority carrier continuity equation by

$$(9) \qquad \frac{\partial p}{\partial t} = - \frac{p - p_0}{\tau_p} - \mu_p \mathcal{E} \frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} \qquad \text{(N-type)}$$

or

$$(10) \qquad \frac{\partial n}{\partial t} = - \frac{n - n_0}{\tau_n} + \mu_n \mathcal{E} \frac{\partial n}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} \qquad \text{(P-type)} \quad .$$

The condition that $\frac{\partial \mathcal{E}}{\partial x}$ is small inplies that $\mathcal{E}$ is nearly constant. This suggests that (9) and (10) are essentially linear differential equations. The particular situation in which

$$(11) \qquad D_p \frac{\partial^2 p}{\partial x^2} >> \mu_p \mathcal{E} \frac{\partial p}{\partial x} \qquad \text{(N-material)}$$

is referred to as the diffusion model approximation. It is of great importance in device analysis. Its implications may be studied as follows: The statement (11) means $J_{p_{DIFFUSION}} >> J_{p_{DRIFT}}$. If this condition is satisfied (9) becomes

$$(9a) \quad \frac{\partial p}{\partial t} \approx \frac{p - p_0}{\tau_p} + D_p \frac{\partial^2 p}{\partial x^2} \qquad \text{(N-material)}$$

or, in the steady state

$$(9b) \quad 0 \approx \frac{p - p_0}{\tau_p} + D_p \frac{d^2}{dx^2} (p-p_0)$$

where we assume constant doping: $n_0 = N_D$, $p_0 = \frac{n_i^2}{N_D}$. The solution of this equation is of the form

$$(12) \quad p - p_0 = A\varepsilon^{x/L_p} + B\varepsilon^{-x/L_p} \qquad L_p = \sqrt{D_p \tau_p} \ .$$

Here L_p is the minority carrier diffusion length, which we defined earlier as the average distance which a minority carrier can diffuse before it recombines. If we assume that the N-type semiconductor extends from $0 \leq x \leq \infty$ (12) reduces to

$$(13) \quad p - p_0 = (p-p_0)\Big|_{x=0} \varepsilon^{-x/L_p} \ .$$

The diffusion current which is associated with this distribution is

$$J_p = -qD_p \frac{dp}{dx}$$

or

$$(13) \quad J_p = \frac{qD_p}{L_p} (p-p_0)\Big|_{x=0} \varepsilon^{-x/L_p} \ .$$

Since we are assuming steady state conditions

$$(5) \quad \overline{\nabla} \cdot \overline{J} = 0 \qquad J = \text{constant}$$

hence $\qquad J_N = J - J_p$.

We define the injection efficiency by

$$(14)\quad \gamma = \frac{J_P(x)}{J} = \frac{J_P(0)}{J}\,\varepsilon^{-x/L_p} = \gamma(0)\varepsilon^{-x/L_p} .$$

The two currents may therefore be written as

$$J_P(x) = \gamma(x)J$$

$$(14a)\quad J_N(x) = \left(1 - \gamma(x)\right)J .$$

The electron current has two components: drift and diffusion. Hence:

$$J_{nDF} = qD_n \frac{dn}{dx} = qD_n \frac{d}{dx}(n-n_0) = qD_n \frac{d}{dx}(p-p_0)$$

$$= \frac{D_n}{D_p}\, qD_p \frac{d(p-p_0)}{dx}$$

$$(15)\quad J_{nDF} = -bJ_P(0)\varepsilon^{-x/L_p} \qquad b = \frac{\mu_n}{\mu_p} .$$

The electron drift current is

$$J_{nDR} = q\mu_n n E$$

but for small injection $n = n_0 + \Delta n \approx n_0$. Hence

$$J_{NDR} \approx q\mu_n n_0 E$$

$$(16)\quad J_{NDR} = \sigma E \qquad \sigma = q\mu_n n_0 + q\mu_p p_0 \approx q\mu_n n_0 .$$

The total electron current becomes

$$J_N = J_{NDF} + J_{NDR}$$

$$(17)\quad J_N = \sigma E - bJ_P(0)\varepsilon^{-x/L_p}$$

but $$= \sigma E - b\gamma(0)J\varepsilon^{-x/L_p}$$

$$(14a)\quad J_N = \left(1 - \gamma(x)\right)J$$

$$= \left(1 - \gamma(x)\varepsilon^{-x/L_p}\right)J \ .$$

Solving for the field yields

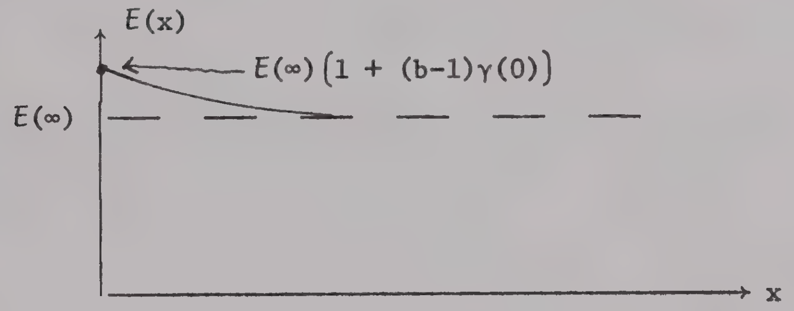
$$(18)\quad E = \frac{J}{\sigma}\left|1 - (1-b)\gamma(0)\varepsilon^{-x/L_p}\right| \ .$$

The field at very large values of x is $\frac{J}{\sigma} = E(\infty)$. Thus

$$(18a)\quad E(\infty) = \frac{J}{\sigma}$$

$$E(x) = E(\infty)\left|1 + (b-1)\gamma(0)\varepsilon^{-x/L_p}\right|$$

for silicon $b \simeq 3$. The resulting field appears as follows:



Our analysis is valid only if the minority carrier drift current is small: Eq. (11). We may check this as follows

$$J_{pDR} = q\mu_p(p_0+\Delta p)E \ .$$

The worst case drift current appears at x = 0.

$$J_{pDR}(0) = q\mu_p\left(p_0 - p(0)\right)E(0) \ .$$

This current has to be smaller than the hole diffusion current

$$q\mu_p\left(p_0 + \Delta p(0)\right)E(0) << \frac{qD_p\Delta p(0)}{L_p} \ .$$

If $\Delta p > > p_0$ this yields

$$\mu_p \mathcal{E}(0) < < \frac{D_p}{L_p} = \frac{kT}{q} \frac{\mu_p}{L_p}$$

or

$$L_p < < \frac{kT}{q} \frac{1}{\mathcal{E}(0)}$$

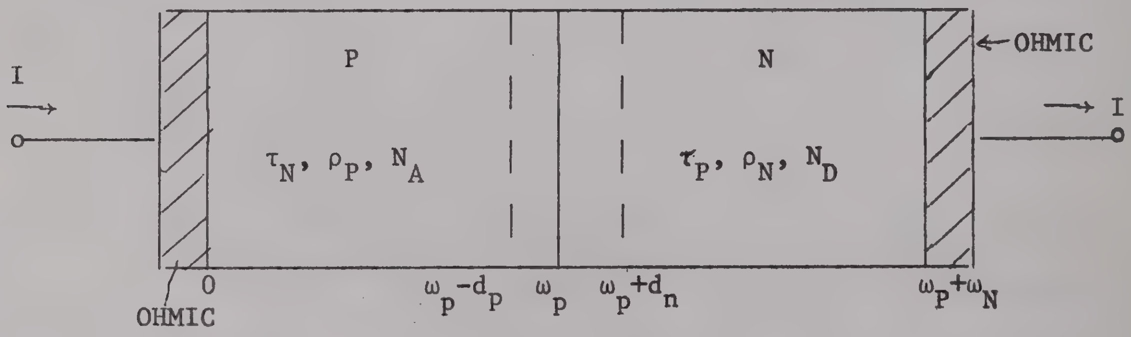
or

$$(19) \quad \mathcal{E}(0) < < \frac{kT}{qL_p} \quad .$$

Thus, as long as Eq. (19) is satisfied the minority carrier drift current may be neglected. Consequently we may calculate the minority carrier distribution from the diffusion equation (9a). We conclude that the diffusion model is valid as long as $\mathcal{E} < \frac{kT}{q} \frac{1}{L_p}$.

Diffusion Model of the Diode

Suppose we apply our ideas to the junction diode.



Examination of the problem suggests that we have three regions:

I. $0 < x < \omega_p - d_p$ (P-Bulk)
II. $\omega_p - d_p < x < \omega_p + d_n$ (Space Charge Region)
III. $\omega_p + d_n < x < \omega_P + \omega_N$ (N-Bulk) .

If we restrict the current I to low enough values the field in the "neutral" bulk will be small enough to satisfy $\mathcal{E}_I < \frac{kT}{qL_N}$ and $\mathcal{E}_{III} < \frac{kT}{qL_P}$. In that case the diffusion model applied to both bulk regions. The space charge region always will have a large field. The diffusion model will not apply.

Region I. $0 < x < w_p - d_p$

The diffusion model equation is

$$(1) \qquad 0 = -\frac{\Delta n}{\tau_n} + D_n \frac{d^2\Delta n}{dx^2} ,$$

its solution is

$$(2) \qquad \Delta n = A_1 \varepsilon^{-x/L_N} + B_1 \varepsilon^{x/L_N} \qquad\qquad L_N = \sqrt{D_n \tau_n}$$

since the contact at x = 0 is ohmic

$$(3) \qquad \Delta n(0) = 0 .$$

At the space charge edge we write

$$(4) \qquad \Delta n(w_p - d_p) = \Delta n(W_p)$$

then

$$(5) \qquad \begin{bmatrix} \Delta n(W_p) \\ 0 \end{bmatrix} = \begin{bmatrix} \varepsilon^{-W_p/L_N} & \varepsilon^{W_p/L_N} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ B_1 \end{bmatrix}$$

or

$$(6) \qquad \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} = \frac{1}{\varepsilon^{W_p/L_N} - \varepsilon^{-W_p/L_N}} \begin{bmatrix} -1 & \varepsilon^{W_p/L_N} \\ 1 & -\varepsilon^{-W_p/L_N} \end{bmatrix} \begin{bmatrix} \Delta n(W_p) \\ 0 \end{bmatrix} .$$

Substituting into (2) yields:

$$(7) \qquad \Delta n(x) = \frac{\sinh x/L_N}{\sinh \dfrac{w_p - d_p}{L_N}} \Delta n(w_p - d_p) .$$

We note that (1) is a linear differential equation. (7) states that $\Delta n(x)$ is linearly related to $\Delta n(w_p\text{-dp})$. The minority carrier current is

and

$$(13)\quad J_p = \frac{qD_p}{L_p}\,\frac{\cosh \dfrac{w_p + w_N - x}{L_p}}{\sinh \dfrac{w_N - d_n}{L_p}}\,\Delta p(w_p+d_n)\ .$$

Region II. $w_p - d_p < x < w_p + d_n$

This is the space-charge region. Surely we cannot describe it by the diffusion equation which requires a small field. But a space-charge region, if it is a depletion region, has nearly no mobile carriers. This in turn would imply no recombination. But if there is no recombination and if we neglect generation also then if J_N enters this region J_N will also leave it. In other words

$$(14)\quad J_N(w_p-d_p) \approx J_N(w_p+d_n)$$

$$(15)\quad J_P(w_p-d_p) \approx J_P(w_p+d_n)$$

but the total current is $J = J_N(x) + J_P(x)$. Hence:

$$(16)\quad J = J_N(w_p-d_p) + J_P(w_p+d_n)\ ,$$

substituting from (13) and (8) yields

$$(17)\quad J = \left(\frac{qD_n}{L_N}\coth\frac{w_p - d_p}{L_N}\right)\Delta n(w_p-d_p) + \left(\frac{qD_p}{L_P}\coth\frac{w_N - d_n}{L_P}\right)\Delta p(w_p+d_n)\ .$$

This equation suggests that the current is linearly related to the excess densities near the junction. In order to make (17) useful we must next relate the excess densities to the applied voltage which causes the current density J to flow.

Voltage-Excess Density Relationships:

Consider the equation

$$(1) \qquad J_N = q\mu_n n\mathcal{E} + qD_n \frac{dn}{dx}$$

or

$$(1a) \qquad \frac{dn}{dx} = \frac{J_n}{qD_n} - \frac{q}{kT} n\mathcal{E} \qquad\qquad \left(\frac{q}{kT} = \frac{\mu_n}{D_n}\right)$$

The field is related to the potential by $\mathcal{E} = -\frac{dv}{dx}$. In order to integrate (1a) consider

$$\frac{d}{dx}\left[n(x)\varepsilon^{-\frac{qv(x)}{kT}}\right] = \frac{dn}{dx}\varepsilon^{-\frac{qv(x)}{kT}} - \frac{q}{kT}\frac{dv}{dx}n\varepsilon^{-\frac{qv(x)}{kT}}$$

$$= \left(\frac{dn}{dx} + \frac{q}{kT}n\mathcal{E}\right)\varepsilon^{-\frac{qv(x)}{kT}}$$

but then

$$(1b) \qquad \left(\frac{dn}{dx} + \frac{q}{kT}n\mathcal{E}\right)\varepsilon^{-\frac{qv(x)}{kT}} = \frac{J_N}{qD_n}\varepsilon^{-\frac{qv(x)}{kT}}$$

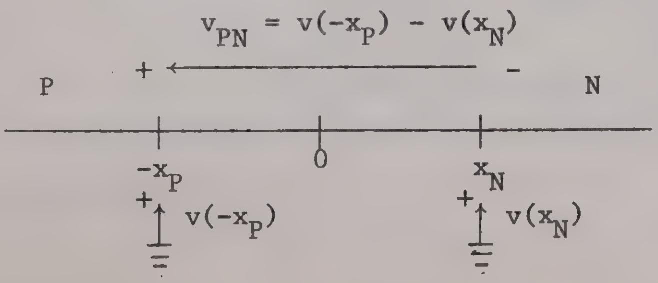
$$= \frac{d}{dx}\left[n(x)\varepsilon^{-\frac{qv(x)}{kT}}\right] .$$

We may use a similar procedure on the hole current equation. This yields:

$$(2) \qquad \frac{d}{dx}\, n(x)\varepsilon^{-\frac{qv(x)}{kT}} = \frac{J_N}{qD_n}\varepsilon^{-\frac{qv(x)}{kT}}$$

$$(3) \qquad \frac{d}{dx}\, p(x)\varepsilon^{\frac{qv(x)}{kT}} = -\frac{J_P}{qD_p}\varepsilon^{\frac{qv(x)}{kT}} .$$

Let us apply these two equations to the space charge region. For convenience we shift the coordinate system:



Integrating (2) and (3) yields

$$n(x_N)\varepsilon^{-\frac{qv(x_N)}{kT}} - n(-x_P)\varepsilon^{-\frac{qv(-x_P)}{kT}} = \int_{-x_P}^{x_N} \frac{J_N}{qD_n}\,\varepsilon^{-\frac{qv(x)}{kT}}\,dx$$

$$p(x_N)\varepsilon^{\frac{qv(x_N)}{kT}} - p(-x_P)\varepsilon^{\frac{qv(-x_P)}{kT}} = -\int_{-x_P}^{x_N} \frac{J_P}{qD_p}\,\varepsilon^{\frac{qv(x)}{kT}}\,dx\ .$$

We may solve these two equations for the minority carrier densities.

$$(4)\quad n(-x_P) = n(x_N)\varepsilon^{\frac{q}{kT}\left(v(-x_P)-v(x_N)\right)} - \varepsilon^{\frac{q}{kT}v(-x_P)}\int_{-x_P}^{x_N} \frac{J_N}{qD_p}\,\varepsilon^{-\frac{qv(x)}{kT}}\,dx$$

$$(5)\quad p(x_N) = p(-x_P)\varepsilon^{\frac{q}{kT}\left(v(-x_P)-v(x_N)\right)} - \varepsilon^{-\frac{q}{kT}v(-x_N)}\int_{-x_P}^{x_N} \frac{J_P}{qD_p}\,\varepsilon^{\frac{qv(x)}{kT}}\,dx\ .$$

Suppose now that we neglect the two integrals. Then we have

$$(4a)\quad n(-x_P) = n(x_N)\varepsilon^{\frac{q}{kT}\left(v(-x_P)-v(x_N)\right)}$$

$$(5a)\quad p(x_N) = p(-x_P)\varepsilon^{\frac{q}{kT}\left(v(-x_P)-v(x_N)\right)}\ .$$

The junction potential

$$v_J = v_{PN} = v(-x_P) - v(x_N)$$

contains the built-in potential. It is related to the external potential: v_{JEX} by

$$(6)\quad v_{JEX} = v_J - v_0 = v(-x_P) - \left(v(-x_{P0})-v(x_{N0})\right)\ .$$

We may write Eq. (4a) at equilibrium as:

$$(4b)\quad 1 = \frac{n_0(-x_{P0})}{n_0(x_{N0})}\,\varepsilon^{-\frac{q}{kT}\left(v(-x_{0P})-v(x_{N0})\right)}$$

$$(5b)\quad 1 = \frac{p_0(x_{N0})}{p_0(-x_{P0})}\,\varepsilon^{-\frac{q}{kT}\left(v(-x_{0P})-v(x_{N0})\right)}\ .$$

Combining (4b), (6) and (4a) and (5b), (6) and (5a) yields

$$(4c) \quad n(-x_P) = n_0(-x_{P0}) \frac{n(x_N)}{n_0(x_{N0})} \varepsilon^{\frac{qv_{JEX}}{kT}}$$

$$(5c) \quad p(x_N) = p_0(x_{N0}) \frac{p(-x_P)}{p_0(-x_{P0})} \varepsilon^{\frac{qv_{JEX}}{kT}} .$$

Let us next assume that we have charge neutrality at the edges of the space charge region.

$$(7) \quad p(-x_P) = N_A(-x_P) + n(-x_P)$$

$$(8) \quad n(x_N) = N_D(x_N) + p(x_N)$$

Eliminating the majority carrier term from (4c) and (5c) yields

$$(4d) \quad n(-x_P) - \left(\frac{n_0(-x_{P0})}{n_0(x_{N0})} \varepsilon^{\frac{qv_{JEX}}{kT}}\right) p(x_N) = \frac{n_0(-x_{P0})}{n_0(x_{N0})} N_D(x_N) \varepsilon^{\frac{qv_{JEX}}{kT}}$$

$$(5d) \quad - \left(\frac{p_0(x_{N0})}{p_0(-x_{P0})} \varepsilon^{\frac{qv_{JEX}}{kT}}\right) n(-x_P) + p(x_N) = \frac{p_0(x_{N0})}{p_0(-x_{P0})} N_A(-x_P) \varepsilon^{\frac{qv_{JEX}}{kT}} .$$

This is a set of equations for the unknowns $n(-x_P)$ and $p(x_N)$. If the doping is reasonable:

$$n_0(x_{N0}) \approx N_D(x_{N0}) \qquad p_0(-x_{P0}) \approx N_A(-x_{P0})$$

$$\frac{n_0(-x_{P0})}{n_0(x_{N0})} = \frac{n_i^2}{n_0(x_{N0}) p_0(-x_{P0})} \approx \frac{n_i^2}{N_D(x_{N0}) N_A(-x_{P0})} ,$$

these equations are used to derive (9):

$$(9) \quad n(-x_P) = \frac{\left[N_D(x_N) + N_A(-x_P) \frac{n_i^2}{N_D(x_{N0}) N_A(-x_{P0})} \varepsilon^{\frac{qv_{JEX}}{kT}}\right] n_i^2 \varepsilon^{\frac{qv_{JEX}}{kT}}}{\left[1 - \left(\frac{n_i^2}{N_A(-x_{P0}) N_D(x_{N0})} \varepsilon^{\frac{qv_{JEX}}{kT}}\right)^2\right] N_A(-x_{P0}) N_D(x_{N0})}$$

The equation for $p(x_N)$ is of the same form as Eq. (9). It may be obtained by subscript exchange A → D, D → A.

This equation may be approximated for special cases. Let us assume that

(a) the P and N material are uniformly doped. Then

$$N_D(x_N) = N_D(x_{N0}) = N_D$$

$$N_A(-x_P) = N_A(-x_{P0}) = N_A$$

(b) $$1 >> \frac{n_i^2}{N_A N_D} \varepsilon^{\frac{qv_{JEX}}{kT}} ,$$

this implies that

$$\frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} >> v_{JEX}$$

or $$v_0 >> v_{JEX}$$

The external junction potential is much less than the built-in potential. Then

(9a) $$n(-x_P) \approx \left(\frac{N_D}{N_A} + \frac{n_i}{N_D N_A} \varepsilon^{\frac{qv_{JEX}}{kT}}\right) \frac{n_i^2}{N_D} \varepsilon^{\frac{qv_{JEX}}{kT}} .$$

Suppose now that we make a third assumption:

(c) $$2 \frac{kT}{q} \ln \left(\frac{N_D}{n_i}\right) >> v_{JEX} .$$

Then

(9b) $$n(-x_P) \approx n_{P0} \varepsilon^{\frac{qv_{JEX}}{kT}} .$$

This equation is subject to four conditions:

(1) The integrals in (4) and (5) are small. This will be true if the total depletion layer width is small.

(2) The doping on either side of the junction is uniform.

(3) The applied potential is less than the built-in potential.

(4) The applied potential is less than $\frac{2kT}{q} \ln \frac{N_D}{n_i}$.

If we also use Eq. (9c),

$$(9c) \quad p(x_N) = p_{n0}\varepsilon^{\frac{qv_{JEX}}{kT}} ,$$

the conditions are further restricted by

$$(5) \quad v_{JEX} < < 2 \frac{kT}{q} \ln \frac{N_A}{n_i} .$$

The five restrictions apply to the use of (9c) and (9b). Only the first restriction applies to (9). It is therefore conceivable to use Eq. (9) and Eq. (9c) for a particular situation. (P^+-N Junction.) We will assume here that we have reasonable doping and small applied voltages. Then (9b) and (9c) used in (17) yields

$$(10) \quad J = \left(\frac{qD_n}{L_n} n_{P0} \coth \frac{W_P - d_P}{L_N} + \frac{qD_p p_{n0}}{L_P} \coth \frac{W_N - d_n}{L_P}\right)\left(\varepsilon^{\frac{qv_{JEX}}{kT}} - 1\right) .$$

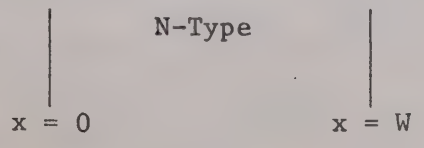
This is the characteristic equation for the diode. Note that if we have long bulk:

$$(10a) \; J \approx \left(\frac{qD_n}{L_n} n_{P0} + \frac{qD_p}{L_p} p_{n0}\right)\left(\varepsilon^{\frac{qv_{JEX}}{kT}} - 1\right) = J_s\left(\varepsilon^{\frac{qv_{JEX}}{kT}} - 1\right) .$$

Here J_s is the saturation current for the device. It does not include the sapce charge generated current.

Transistor Base:

Let us now apply our ideas of the two preceding sections to a piece of uniformly doped material.



Let us apply $\Delta p(0)$ excess holes to the x - 0 boundary. At $x = W$ we assume to have a contact which removes all holes: $\Delta p(W) = -p_{n0}$. We also assume that the diffusion model holds: $E < \frac{kT}{qL_p}$. The solution for the minority carrier density is obtained as follows:

$$(1) \qquad 0 = -\frac{\Delta p}{\tau_p} + D_p \frac{d^2 \Delta p}{dx^2}$$

$$(2) \qquad p = A\varepsilon^{x/L_p} + B\varepsilon^{-x/L_p} \; ;$$

substituting the boundary conditions yields:

$$(3) \qquad \begin{bmatrix} \Delta p(0) \\ -p_{n0} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \varepsilon^{W/L_p} & \varepsilon^{-W/L_p} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}$$

$$(4) \qquad \begin{bmatrix} A \\ B \end{bmatrix} = \frac{2}{\sinh \frac{W_i}{L_p}} \begin{bmatrix} -\varepsilon^{-W/L_p} & 1 \\ \varepsilon^{W/L_p} & -1 \end{bmatrix} \begin{bmatrix} \Delta p(0) \\ -p_{n0} \end{bmatrix}$$

or

$$(5) \qquad \Delta p = \frac{\Delta p(0) \sinh \frac{W - x}{L_p} - p_{n0} \sinh \frac{x}{L_p}}{\sinh \frac{W}{L_p}} \; .$$

The minority carrier current is

$$J_p = -qD_p \frac{d\Delta p}{dx}$$

or

$$(6) \qquad J_p = \frac{qD_p}{L_p} \frac{\Delta p(0) \cosh \frac{W - x}{L_p} + p_{n0} \cosh \frac{x}{L_p}}{\sinh \frac{W}{L_p}} \; .$$

Evaluation at the two boundaries yields:

$$(7) \qquad J_p(0) = \frac{qD_p}{L_p} \Delta p(0) \coth \frac{W}{L_p} + \frac{qD_p}{L_p} p_{n0} \operatorname{csch} \frac{W}{L_p}$$

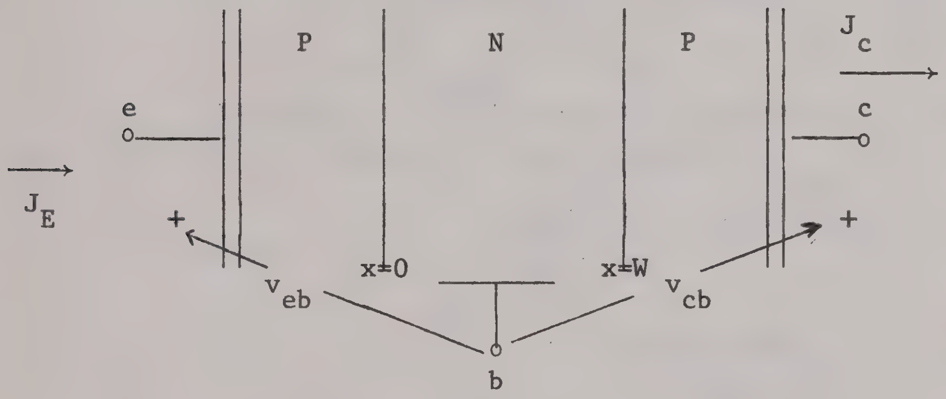
$$(8) \qquad J_p(W) = \frac{qD_p}{L_p} \Delta p(0) \operatorname{csch} \frac{W}{L_p} + \frac{qD_p}{L_p} p_{n0} \coth \frac{W}{L_p} \; .$$

If we make the width of the N-material very small: $W << L_p$ these equations reduce to:

$$(7a) \quad J_p(0) = \frac{qD_p}{W} \Delta p(0) + \frac{qD_p}{W} p_{n0}$$

$$(8a) \quad J_p(W) = \frac{qD_p}{W} \Delta p(0) + \frac{qD_p}{W} p_{n0} \; .$$

The two currents are therefore nearly identical. Let us next consider the contact at x = 0. It supplies the excess density $\Delta p(0)$. We provide this by forward biasing a junction as follows:



But then

$$(9) \quad \Delta p(0) = p_n \left(\varepsilon^{\frac{qv_{EB}}{kT}} - 1 \right)$$

and $$J_p(0) \approx J_p(W) = \frac{qD_p}{W} \varepsilon^{\frac{qv_{EB}}{kT}} p_{n0} \; .$$

The contact to the base was established with the metal contact as shown. This contact will not interfere with the N-material if the field in the N-region is small. But a diffusion current requires no field. Thus, if $J_p(0)$ is the total current the contact will not affect our analysis. We may force the condition $J_p(0) \approx J_E$ by designing the P contact correctly. The excess electron concentration at x = 0 is

$$(10) \quad \Delta n(0) = n_{p0} \left(\varepsilon^{\frac{qv_{EB}}{kT}} - 1 \right) \; .$$

If the P-material is very long:

$$(11) \quad \Delta n(x) = \Delta n(0)\varepsilon^{x/L_N}$$

and
$$J_N = \frac{qD_n}{L_N} \varepsilon^{x/L_N} \Delta n(0)$$

hence

$$(11a) \quad J_N(0) = \frac{qD_N}{L_N} n_{p0} \left(\varepsilon^{\frac{qv_{EB}}{kT}} - 1 \right) .$$

The total current at x = 0 becomes

$$(12) \quad J_E = \frac{qD_n}{L_N} n_{p0} \left(\varepsilon^{\frac{qv_{EB}}{kT}} - 1 \right) + \frac{qD_p}{W} \varepsilon^{\frac{qv_{EB}}{kT}} p_{n0} .$$

The hole injection efficiency at x = 0 for $\varepsilon^{\frac{qv_{EB}}{kT}} >> 1$ becomes

$$(13) \quad \gamma_E = \frac{J_P(0)}{J_N(0) + J_P(0)} = \frac{1}{1 + \frac{J_N(0)}{J_P(0)}}$$

$$= \frac{1}{1 + \frac{D_n n_{p0}}{L_N} \frac{W}{D_p p_{n0}}}$$

or

$$(13a) \quad \gamma_E \approx \frac{1}{1 + \frac{W}{L_N} \frac{n_{p0}}{p_{n0}}} .$$

We would like to have this value as close to unity as possible. In order to do this let

$$p_{n0} >> n_{p0}$$

or
$$\frac{n_i^2}{N_D} >> \frac{n_i^2}{N_A}$$

or
$$N_A >> N_D .$$

Thus, if we dope the P-material heavy compared to the N-material nearly all of the current which crosses into the N-material will be a hole current which ends up at x = W and not at the metal contact. We write

$$\text{(12a)}\quad J_E \approx \frac{qD_p}{W} p_{n0}\, \varepsilon^{\frac{qv_{EB}}{kT}} \qquad N_A >> N_D \ .$$

Let us next consider the situation at x = W. Here $\Delta p(W) = -p_{n0}$. We supply a reverse biased junction as shown.

$$\text{(14)}\quad \Delta p(W) = p_{n0}\left(\varepsilon^{\frac{qv_{CB}}{kT}} - 1\right)$$

$$= -p_{n0} \qquad \text{if} \qquad v_{CB} \to -\infty \ .$$

The excess electron density in the P-material is

$$\Delta n(W) = -n_{p0} \qquad \text{if} \qquad v_{CB} \to -\infty$$

$$\Delta n(x) = -n_{p0}\, \varepsilon^{-\frac{x-W}{L_N}} \qquad \text{(long P-bulk)}$$

hence $$J_n = \frac{qD_n}{L_N} n_{p0}\, \varepsilon^{-\frac{x-W}{L_N}} \ .$$

The total current at x = W becomes

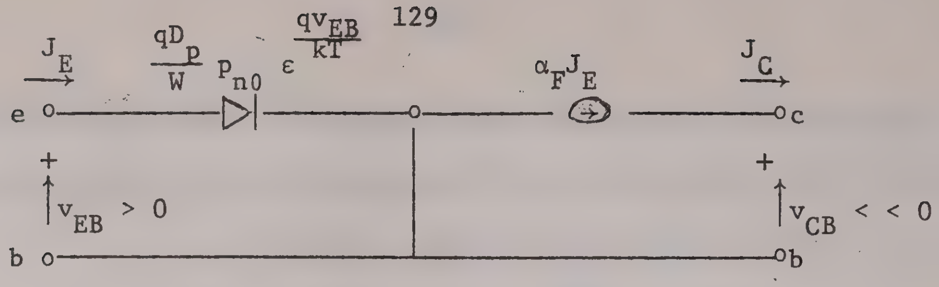
$$J_c = J_P(W) + J_N(W)$$

$$= \frac{qD_p}{W} p_{n0}\, \varepsilon^{\frac{qv_{EB}}{kT}} + \frac{qD_n}{L_N} n_{p0}$$

or

$$\text{(15)}\quad J_c \approx \frac{qD_p}{W} p_{n0}\, \varepsilon^{\frac{qv_{EB}}{kT}} \ .$$

Comparing Eq. (12a) and (15) yields the following equivalent circuit.



Here $\alpha_F = \frac{J_C}{J_E} \simeq 1$ for our idealized structure.

Supplementary Notes for Course 330

The Importance of the Concept of Frequency in the Analysis and Synthesis of Communication Circuits

How Did it Begin and Develop

The first electric communication system was the electric telegraph beginning a little before 1850. As the transmission lines connecting transmitter and receiver became longer and also were put under sea, the received signals grew more distorted so that recognition was difficult or impossible. Figure 1 shows the distortion of a single telegraph dot caused by sending it through a long undersea cable. In answering the question how does this distortion arise? — scientists William Thomson (Lord Kelvin), Heaviside, Poincairé, Pupin, Campbell and others who studied this problem, used theory worked out by the French mathematician, Fourier, in his study of heat flow. In applying this theory, the train of on-off telegraph pulses is replaced by its equivalent, a sum of sinusoidal waves. A sinusoidal wave is the simplest kind of oscillation found in nature. As we have already seen in our study of Fourier series, the rates of oscillation, or frequencies, of the sinusoids in the sum are harmonically related. In the simplest situation these frequencies are harmonics (that is integral multiples) of the rate at which the telegraph pulses are transmitted.

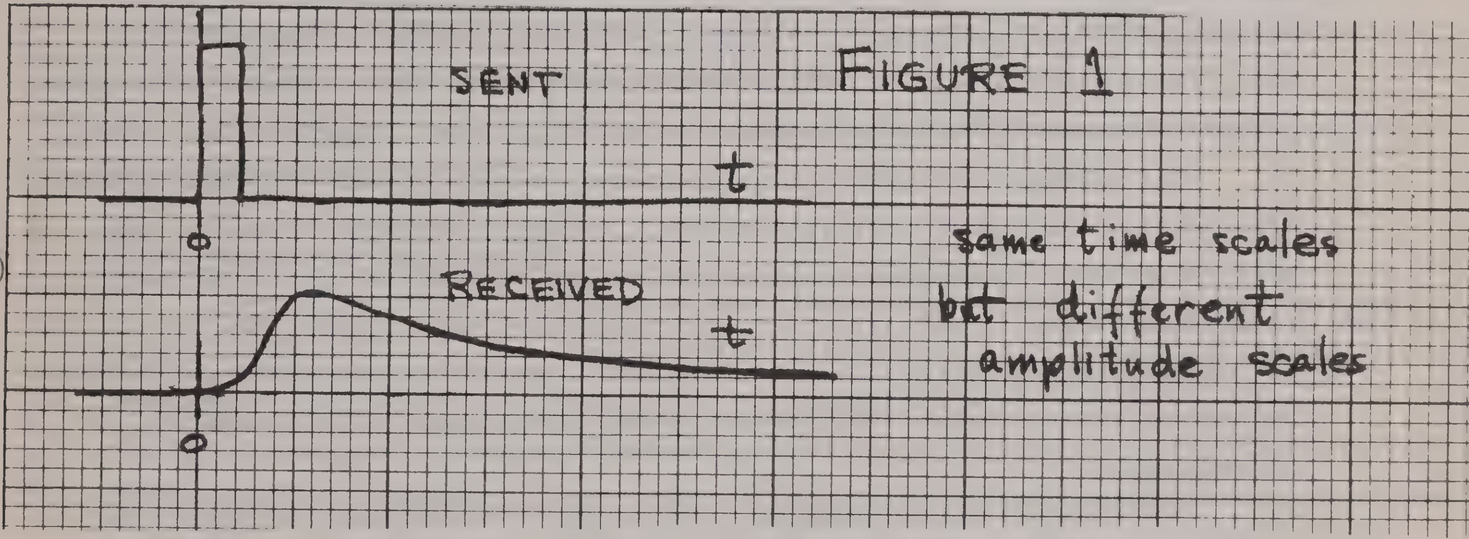


FIGURE 1

With this view of the telegraph signal, the answer to the question above was found. The transmission line did not treat all the frequency components of the signal the same. The amplitudes and phases of some components were changed much more than others in the process of transmission. Hence the sum at the far end of the line did not add up to the waveform as at the transmitter, but to some distortion of it.

This new "frequency" viewpoint provided valuable insight into the behavior of the electrical system. These studies were further stimulated by the invention in 1876 of the telephone, the signal waves of which were considerably more complicated than those of the telegraph. A result of this was that how a given electrical network would carry a given signal could be determined by finding out its response individually to each of a suitable set of the important frequency components of the signal. The Fourier transform methods in which the frequency components are treated as a sum, rather than individually, was not thoroughly developed or widely used until about 25 years ago.

This new idea of representing a signal by a sum of simple sinusoidal signals was demonstrated in the case of sound waves by the great physicist Helmholtz. By combining certain sound resonators (that is an acoustic system which would greatly emphasize a vibration at the single frequency of its resonance), he was able to (partially no doubt) synthesize particular sounds by the particular set of resonators used. He had used these resonators first to analyze various sounds to determine which frequencies were the important ones for that sound.

Not only were our present day concepts and capabilities of time analysis of signals and circuits unavailable then, they were unthinkable in the sense that these modern concepts and capabilities had to wait for the development of the understanding of signals and circuits which we have today. The introduction of Fourier theory and the frequency concept to provide a description of signals and circuits was a great step forward in the development of our present understanding.

But even if in some imagined way the early development had been along concepts of time function alone, our present sophisticated understanding of communication systems could not have been reached. For the concept of frequency is essential to much of our theory and practice. It has been said[†] "Fourier's contribution is absolutely essential to all communication and to communication theory."

Consider how a signal might be described by time properties alone. Could we say that it had no slope higher than a certain one? But there can be many different signals with a highest slope. And how important to the signal is this highest slope? Or consider how a network would be described by its time response alone. We know that the impulse response is sufficient to characterize a network, but many important details are obscured in this presentation.

[†] J. R. Pierce, Symbols, Signals and Noise, p. 30.

The facts are that a transmission system can by physically characterized much more precisely in the frequency domain than in the time domain. And this is extremely important in a system consisting of many repeated links such as are in a transcontinental telephone line.

Actually both time and frequency descriptions are important in our modern understanding and use of signals and systems. One of the important uses of Fourier theory is the transforming back and forth between time and frequency descriptions. It should be realized that there are always two aspects of our problem — 1. signal description and 2. system or network description — whether we are concerned with telegraph, speech, control, data, or picture signals and, whether we are sending them over a two wire line, a waveguide, free space, a simple filter, or a short connection in an integrated circuit.

It seems obvious that when the waveform of a signal is to be kept within certain bounds or has some special use as in telemetry, telegraph, digital transmission, that a time description is essential.

But let us consider some of the ways in which a frequency description is necessary or very helpful.

Frequency Division Multiplex

One of the first of these was the discovery that different radio stations could transmit through the same medium simultaneously, each using a different wavelength or frequency, although this was foreshadowed by Alexander Graham Bell's concept of "harmonic telegraph". Work on this concept led him to the invention of the telephone. It is hard to imagine that this multiple use of the medium for radio waves by frequency division could have arisen if signals had been considered from the time viewpoint alone.

We know now of course that multiple use of a transmission medium is possible also through time division with advantages in some situations. We know also that it would not be suitable for broadcasting use, but is most effective for point to point transmission, partly because of the problems of synchronizing. Further, suitable technology was not available in the early days.

As more radio stations came to use the transmission medium, the problem of interference with one another arose. In this way the concept of the "frequency space" required for a given signal arose. If the frequency space for signal 1 overlapped the frequency space of signal 2, they interfered with one another even though their carrier frequencies, or wavelengths, had a finite separation. This concern led to a study of the frequency space required for various signals.

Another factor besides the interference between stations, leading to the study of signal bandwidths, is the interference caused by noise (called static in the early days of radio). It was soon realized that the power of the noise which accompanied the signal at every receiver was directly proportional to the frequency bandwidth of the receiver as it was tuned to the carried frequency of the desired signal. Hence this was another factor in the restriction of receiver bandwidths to the minimum required for the particular signal.

Knowledge of these frequency band requirements, which began for radio, was pursued further and used in the many frequency division systems which eventually put many telephone channels and/or telegraph signals on a single transmission medium such as a single pair line, or a cable, or a single radio carrier.

Frequency Description of Signals

In Figure 2 some effects of the amount of frequency space available on speech and music signals are shown, that is, how much frequency space or how wide a frequency band is needed for good transmission is shown. While this data is given in terms of the frequency bandwidth of a transmission system, it is really a frequency description of the signal. In Figure 3 the frequency description of a certain telegraph signal and a certain television signal are shown. The telegraph signal is one in which a certain 5 digit character is repeated over and over, thus giving a line spectrum or frequency description. The spectra for any other 3 on pulses per character would lie within the dashed sinc function curve. If rounded pulses instead of rectangular ones, as shown, are sent, the required frequency space is much less, as shown, and interference is reduced. This important result, worked out in the 1920's by H. Nyquist, now retired from Bell Telephone Laboratories, showed that the upper frequencies in the spectrum of the telegraph signal were not necessary for its correct reception. Thus the problems of speed of signaling and interference for telegraph signals, thought of naturally as a time phenomenon, were solved by Fourier theory in the frequency domain. The frequency domain results translated again by Fourier theory to the time domain became the desirability of rounded pulses instead of square ones.

The frequency description of the television signal in Figure 3 is for the particular picture shown, but in general, whatever the picture, the energy in the signal tends to be concentrated about the harmonics of the line scanning rate. The knowledge of the concentration of the signal energy in this way, made possible the addition of color components of a picture signal into the

KEUFFEL & ESSER CO

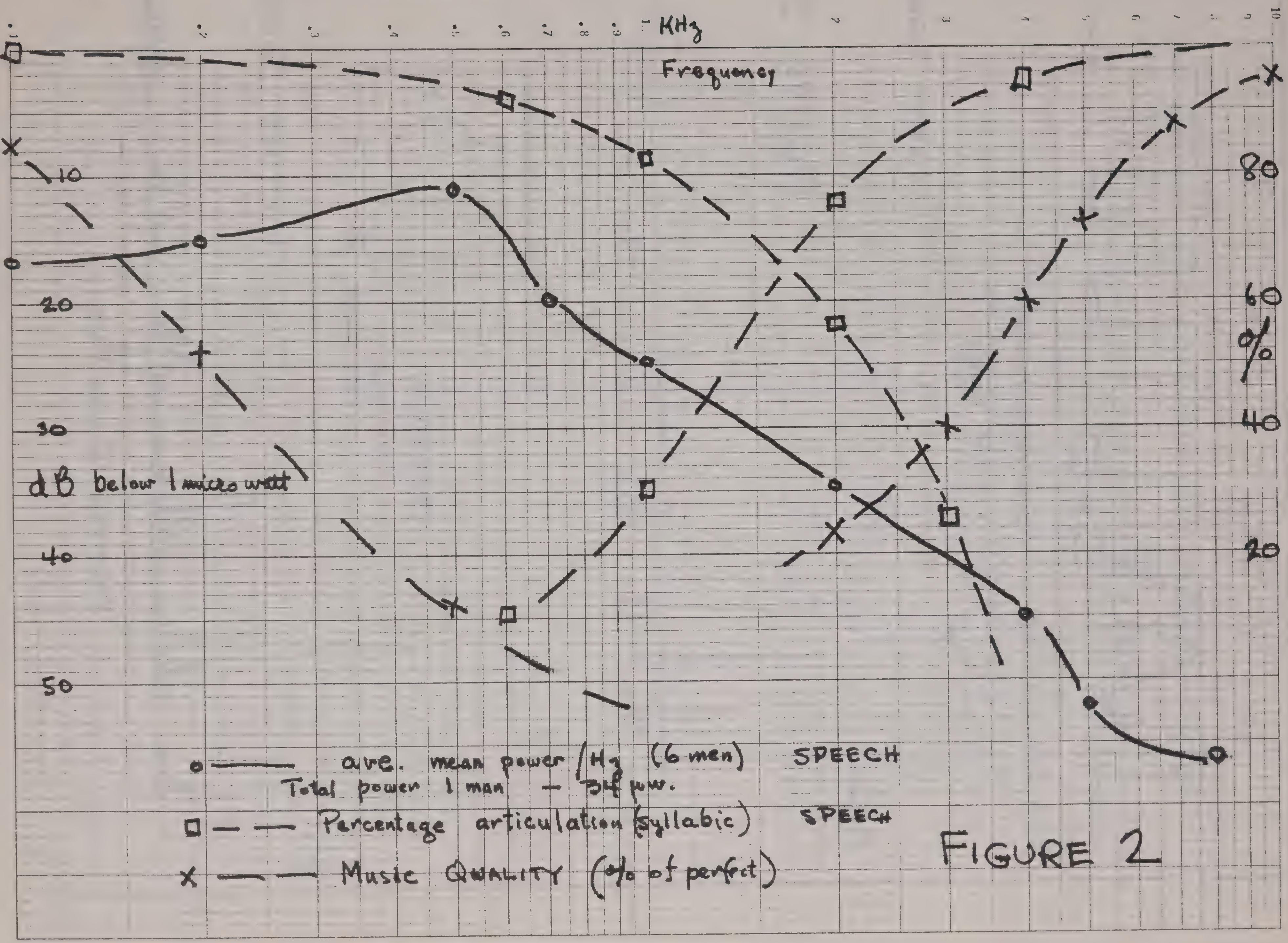


FIGURE 2

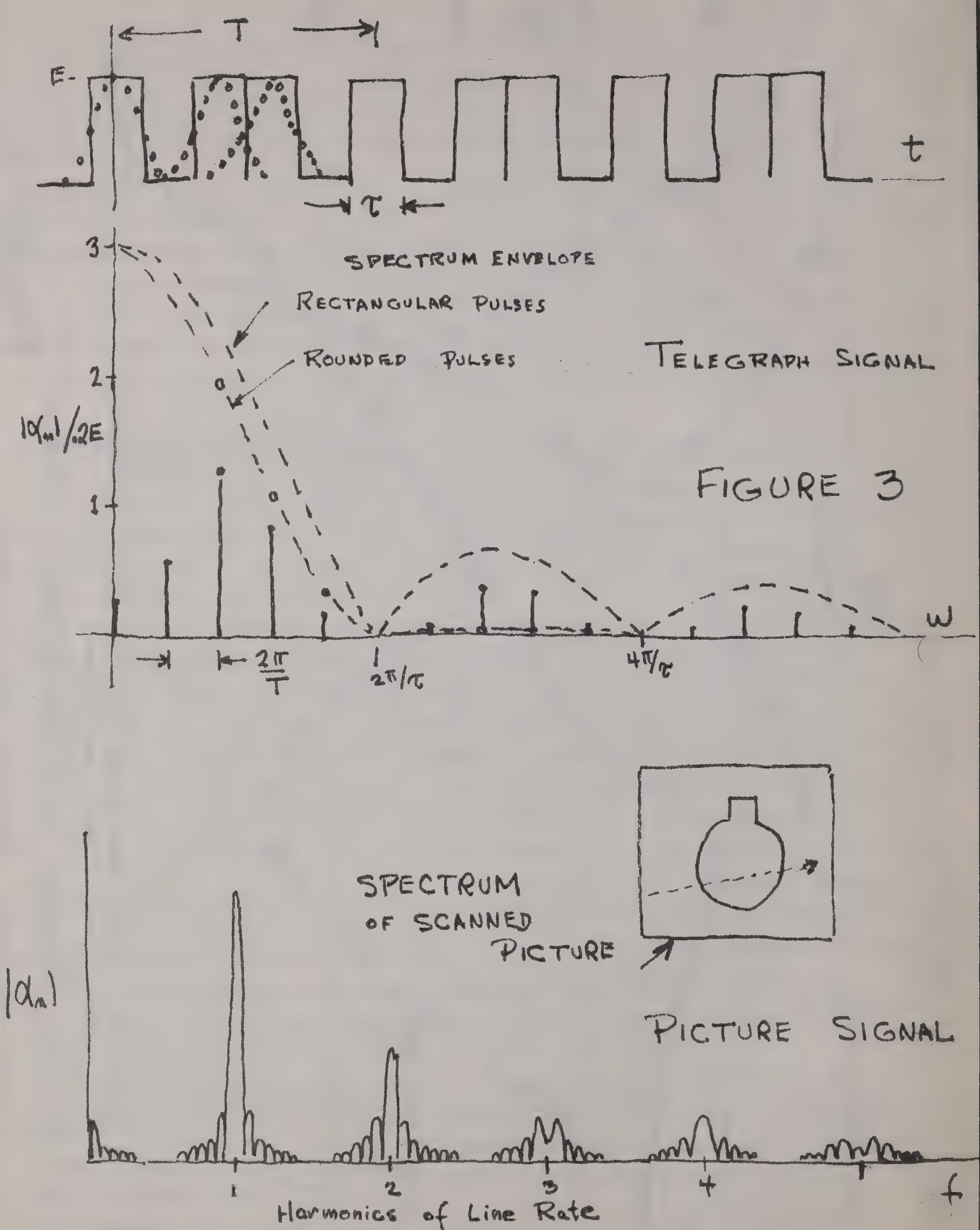


FIGURE 3

7 11/16
10 5/8
ISSUE
U.S.A.
.61)

"empty" frequency spaces with very little additional space required and also was an important factor in the compatibility of color and black and white reception.

If we think instead of the time description of a signal, say the waveform of the sound made in speaking, one of the most outstanding properties is the greatest sharpness of the waveform, or its greatest rate of change. But what slower changes and what their relative strengths are cannot be readily found. A little thought will show that Fourier analysis is equivalent to finding the relative strengths in the signal of a particular set of increasingly faster changes, which together constitute the signal.

Relation Between Time and Frequency Description

There is a very interesting general relationship between the time and frequency description of a signal. If we have a train of pulses, each having duration T, then the frequency band W in which most of the energy of the train lies, is related to T by

$$TW = K_1 \quad , \qquad (1)$$

where K_1 is a number of magnitude near 2π when W is in radians per second and T is in seconds. This means that if the pulses in the illustration are narrow, or in general if the time variations are rapid, then the bandwidth of the signal is large, and vice versa.

Frequency Description of Systems

The system (or network, or circuit, or transmission line) is the necessary means for transporting the signal from one place to another and in general will distort the signal in a negligible or significant way. Our discussion began with the problem of how it does this. A solution to this problem was found by a method which has been used ever since and which has been enlarged considerably. This method, as we have seen, is first to resolve the signal into a particular sum of sinusoids. Then in linear systems, which are the great bulk of communication systems, each of the sinusoids acts independently of the others. Hence their progress through the system may be considered, one at a time. A sinusoid is characterized by its amplitude and phase and frequency. Since we consider one signal component, that is one frequency, at a time, we are concerned only with the first two. If a system changes the amplitude or phase of an important signal component, then it will distort the signal.

This is the frequency description of a system — the amounts it changes the amplitude and phase of sinusoids of all frequencies. It is from the reaction

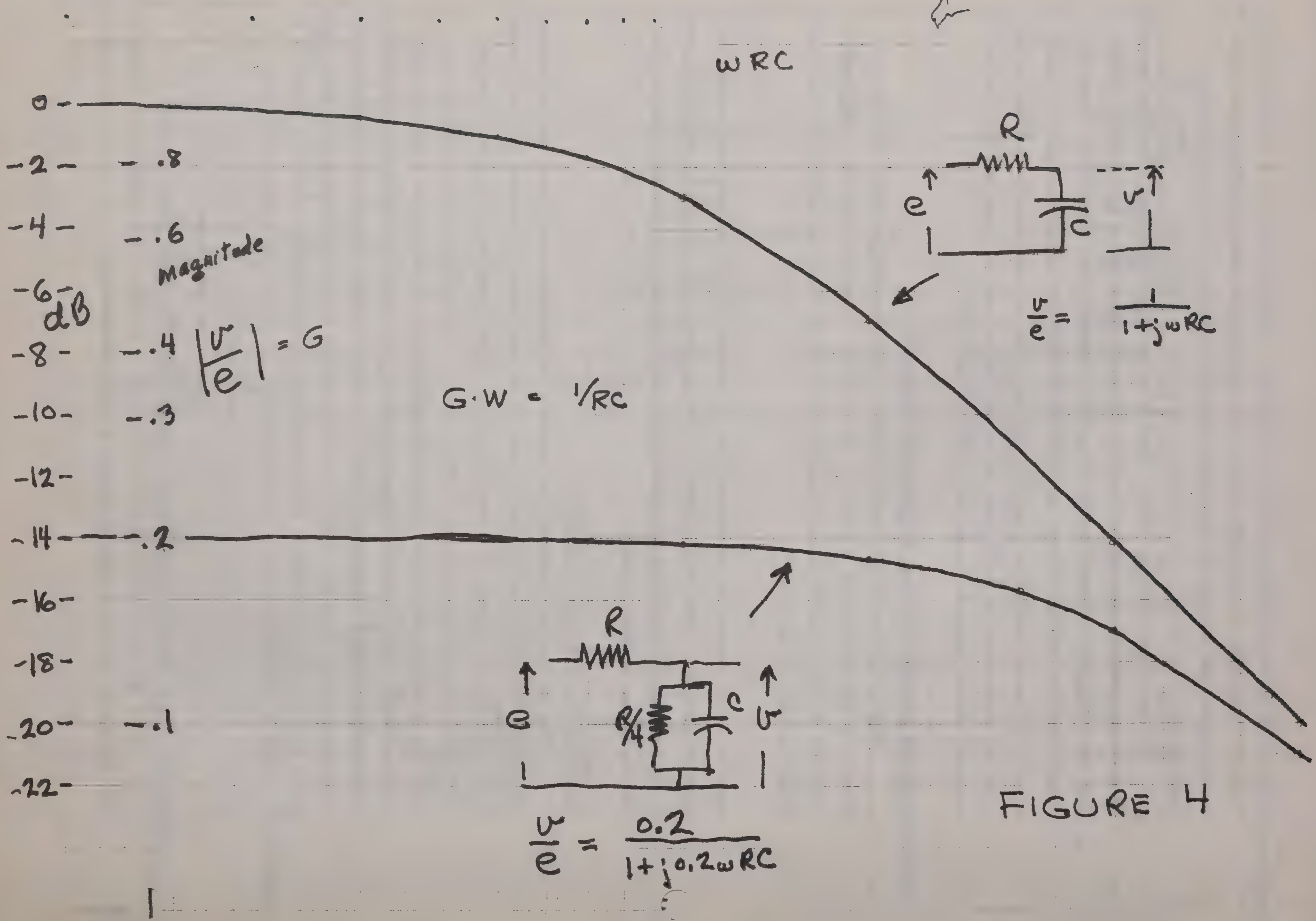

ωRC
0
-2
-4
-6
dB
-8
-10
-12
-14
-16
-18
-20
-22
.8
.6
Magnitude
.4
.3
.2
.1
$\left|\frac{v}{e}\right| = G$
G·W = 1/RC
R
e
C
v
$\frac{v}{e} = \frac{1}{1+j\omega RC}$
R
e
R/4
C
v
$\frac{v}{e} = \frac{0.2}{1+j0.2\omega RC}$

FIGURE 4

of networks to single sinusoids, that the concepts of impedance and admittance, so widely used, have come.

Thus a definite way of studying the performance of systems was found — how the distortion depended on the system and how the system could be changed to reduce the distortion sufficiently.

Now we have seen that a system can be completely described by its impulse response. But it is difficult to relate the details of signal distortion to the details of system structure. This can be done far more readily in the frequency description of systems. Also at present, far more precise measurements of systems may be made in the frequency domain than in the time domain. We have found, or will find, that the frequency description of a system (system function $H(\omega)$) is the Fourier transform of its impulse response $h(t)$. Thus when a very precise time description of a system is needed, it is frequently derived from the frequency description by this relationship, rather than by direct measurement.

Gain Bandwidth Product

There is a relation resembling (1) for systems or networks. It is

$$G \cdot W = K_2 \, , \tag{2}$$

where G is the gain of the system, in voltage ratio for example. For active systems, G can be greater than 1, but for passive systems it must be less than 1. This means that for a given system, we must take a smaller gain if we want a larger frequency band and vice versa. This is illustrated in the curves of Figure 4 for two passive networks. Improvements in devices and systems can sometimes increase the value of the constant K_2. But for a given system we see that the price of a wide band system (large W) is a smaller gain.

This tends to make a wide band system more expensive and more difficult to construct than a narrow one and so we have in this a third reason added to those of least interference with other signals and least added noise, for knowing the minimum frequency band for our signal and then restricting the bandwidth of our system to this.

It is felt by many that these frequency descriptions provide a much more suitable way of dealing with the interactions of signal and system than time descriptions do. Even when time descriptions are used as in the discrete sample representation of signals and in digital filters, the frequency descriptions are usually used in addition. For example, interaction calculations for signal and system are often in the frequency domain after Fourier transforms because the calculation process (like the measuring process) is much simpler and more precise there.

Information Theory

The content of the paragraph before the last one is closely related to the expression of the information capacity of a system. This is

$$\text{Bits/sec.} = W \log_2\left(\frac{P_S}{P_N} + 1\right) \tag{3}$$

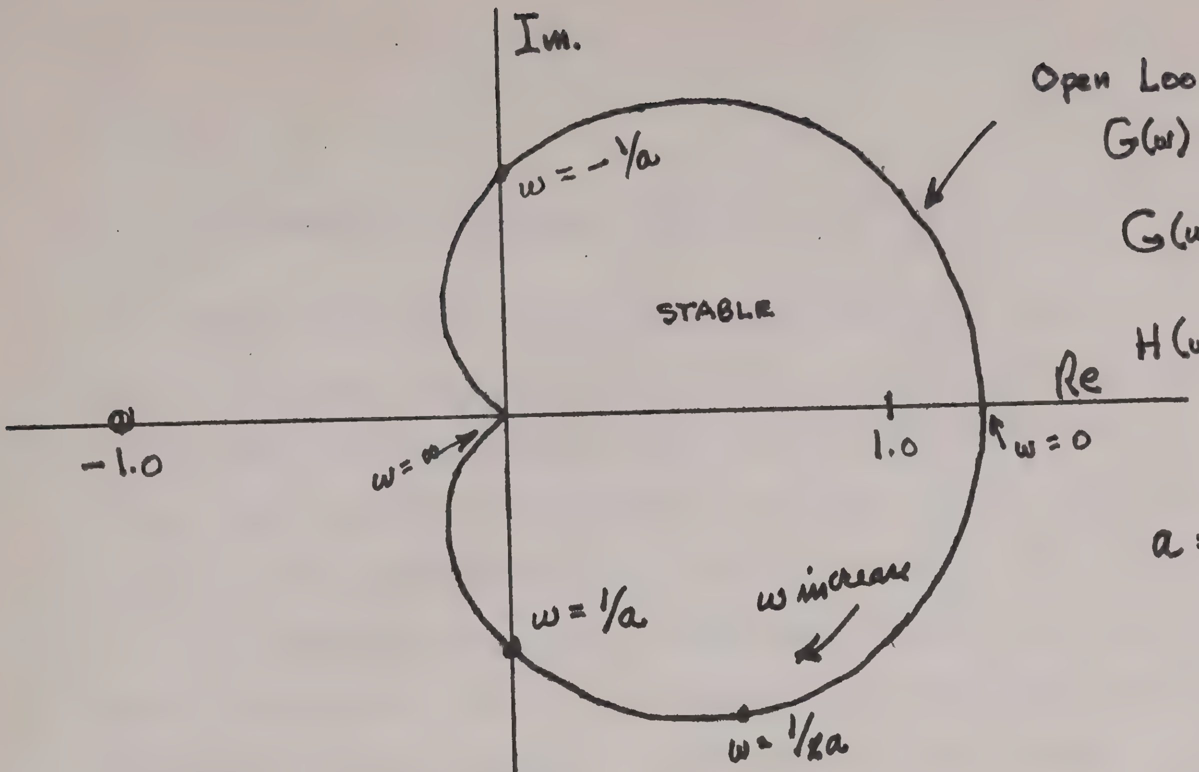
where P_S and P_N are the signal and noise powers coming through the receiving filter of width W.

Stability of Control Circuits

One further example of the importance of the frequency description of systems is found in the Nyquist diagram for determining the stability of feedback or control circuits. H. Nyquist (referred to earlier) in early 1930's transformed early time descriptions of the conditions for stability in a network into the frequency domain. This made possible a graphical distinction between stable and unstable systems not only for idealized systems which could be described mathematically, but for real systems which can only be described accurately by measurements. This is an important point, because a very small detail (that is a small difference between the idealized and real systems) can be the determining factor in questions of stability. Two such Nyquist diagrams are shown in Figure 5.

Conclusion

While it is possible to describe signals and systems in the time domain alone, the use of the frequency domain, brings many advantages both in understanding signals and systems and their interaction and in calculations. Actually both descriptions are valuable, but we would be severely handicapped if we were restricted to time descriptions alone. The frequency description arose first partly for historical reasons and partly because the concepts and technology for it are simpler.

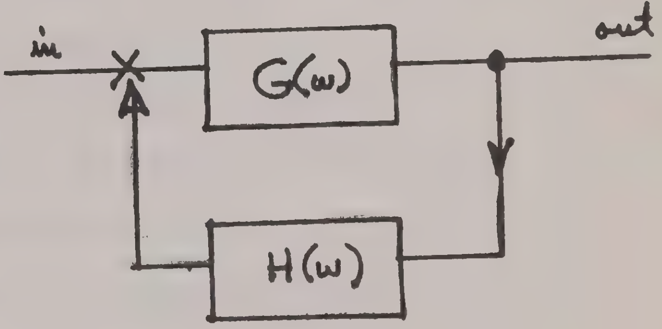


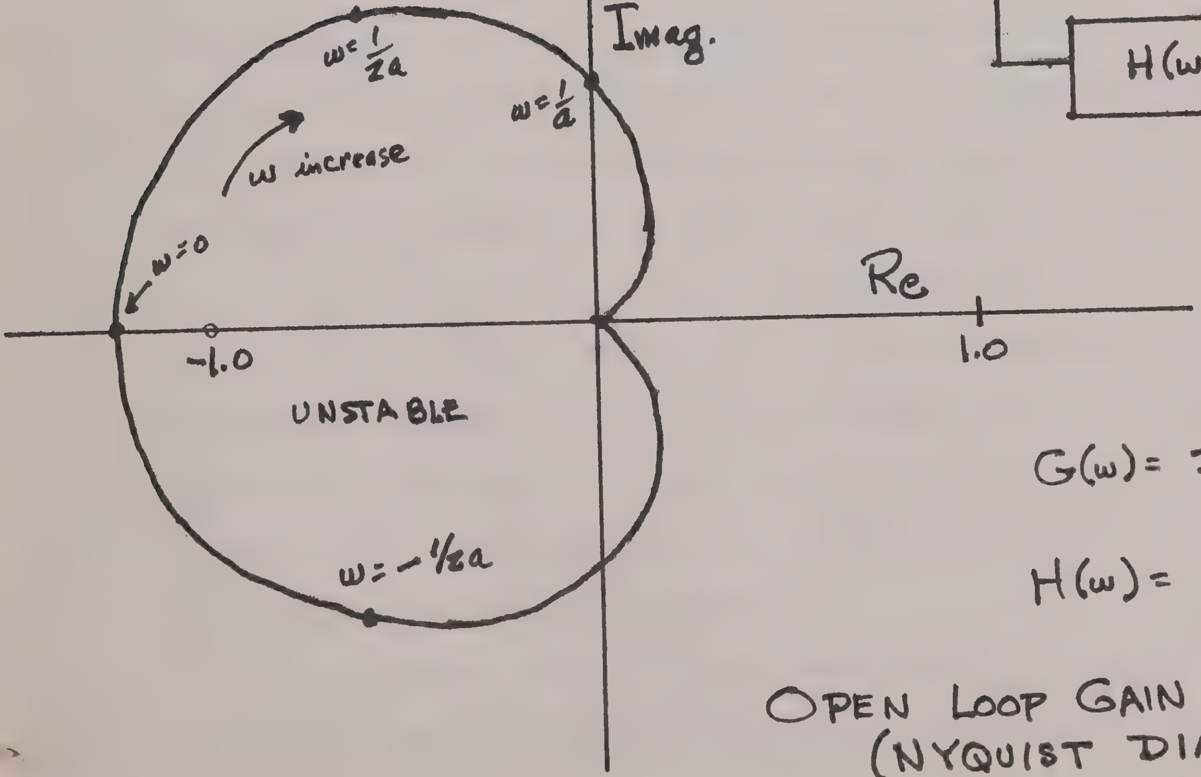
Open Loop Gain
$G(\omega)\,H(\omega)$

$$G(\omega) = \frac{+1.25}{1 + j\omega RC}$$

$$H(\omega) = \frac{1}{1 + j\omega RC}$$

$a = RC$





$$G(\omega) = \frac{-1.25}{1 + j\omega RC}$$

$$H(\omega) = \frac{1}{1 + j\omega RC}$$

OPEN LOOP GAIN STABILITY TEST
(NYQUIST DIAGRAM)

FIGURE 5

1. Power Amplifiers General Comments and Efficiency

While almost all electronic amplifiers are designed to increase the power level of the signal in the circuit, there are certain forms of amplifiers which are designed to deliver substantial power to a transducer or other load. When the main purpose of the amplifier is to deliver power we must be particularly concerned with the voltage and current levels in the output devices as well as the temperature stability and heat dissipation in the devices. Consideration of impedance matching and overall efficiency and distortion are also more important in power amplifiers than in small signal amplifiers.

When we design a power amplifier we must recognize that almost the entire useful active region of the device will be important. A typical power amplifier characteristic and current gain curve are shown in Fig. 1. We must select devices which will have adequately high breakdown voltage and good linearity over the operating region we choose. In the case of transistor power amplifiers, the value of I_{CHi} is most critical. The current gain of the device falls off at higher current densities due to loss of emitter efficiency. We must recognize this when we select the operating load line. Complex emitter base geometries are used to maximize the emitter efficiency while still holding the area of the device to reasonable limits.

Power stages fall into two very general categories. There are those that use transformers to couple power devices to the

load and driver stages and the direct coupled types. Fig. 2 shows typical examples of each type. There are many combinations of CE, CC, CB coupling possible and only a few more common examples will be discussed here.

For a variety of reasons, class B and direct coupled amplifiers are most common.

By the strictest definitions class A means the power stage is biased in such a way that current exists in the devices at all times. This is a relatively linear mode of operation. In class B we should have the devices biased at the edge of the active region so that there is current only during one half of the input cycle. Clearly, a single device would give extreme distortion so a second device is connected to conduct during the "other half" of the cycle. We then have so called "push-pull" connection. Both Fig. 2a and b are push-pull connections in this sense. In fact, we rarely use exact class B but rather some small steady state bias. Thus most transistor amplifiers are "almost class B" or class AB. The reasons for this mode of bias will become clear as we discuss specific cases.

Figure 3 shows bias points and load lines for class A and class B operation. We wich first to consider the efficiency of these classes.

There are two different ways to look at efficiency. One is to define efficiency as

$$\eta = \frac{\text{max power to load}}{\text{max power from supply}}$$

This is useful when we want to know how much of our D.C. power

from the power supply can be converted into load power. In the case of transistor amplifiers however it is also important, if not more important, to consider a merit factor defined by

$$H = \frac{\text{max power to load}}{\text{max power dissipated in transistors}}$$

and to evaluate H for the worst possible signal conditions. This will tell us how hot the devices get at worst when we know the maximum power rating of the amplifier. The efficiency is calculated in the standard way. For class A we have, for sinusoidal signals,

$$P_{LA} = I_L(\text{Avg})\ V_L(\text{Avg})$$

$$= \frac{1}{2}\ \frac{V_{max}}{2}\ \frac{I_{max}}{2} = \frac{1}{8}\ V_{max}\ I_{max}$$

The power delivered to the amplifier by the power supply is

$$P_{CCA} = V_{CC}\ I(\text{Avg})$$

For a transformer coupled load

$$V_{QA} = V_{CC}$$

and

$$P_{CCA} = V_{QA}\ I(\text{Avg}) = \frac{V_{max}}{2}\frac{I_{max}}{2}$$

Then

$$n_A = \frac{P_{LA}}{P_{CCA}} = \frac{V_{max}\ I_{max}/8}{V_{max}\ I_{max}/4} = \frac{1}{2}$$

This means that at full output, the power supply delivers 1 unit of power and 1/2 unit goes to the load while the other 1/2 heats the amplifier.

For class B we can calculate, at full power, the efficiency η_B as follows:

$$P_{LB} = \frac{V_{AB}\ I_{max}}{2}$$

$$P_{CCB} = V_{CC}\ I(Avg) = V_{QB}\ I_{AV}$$

$$= V_{QB}\ I_{max}\ \frac{2}{\pi}$$

Hence

$$\eta_B = \frac{P_{LB}}{P_{CCB}} = \frac{V_{QB}\ I_{max}}{2}\ \ \frac{\pi}{2V_{QB}\ I_{max}} = \frac{\pi}{4}$$

This is 78.5%. Class B is then clearly better than class A since this means that for 1 unit of power from the supply almost 4/5 goes to the load and only 1/5 to heat the amplifier.

It is even more informative if we calculate the power dissipated in the transistors for class B as a function of power out. We can do this readily as follows. For sinusoidal signals,

$$P_{Trans} = 2\ \frac{1}{2\ \pi}\ \int_0^{\pi} p(\omega t) d(\omega t)$$

where $p(\omega t) = (V_Q - IR_L \sin \omega t)\ I \sin \omega t$

Then

$$P_{Trans} = V_Q I\ \frac{2}{\pi} - V_Q\ \frac{I^2}{I_{max}}\ \frac{1}{2} \qquad \text{(A)}$$

We may note that the maximum does not occur at full output but rather setting

$$\frac{dP_{Trans}}{dI} = 0$$

we obtain

$$\frac{2V_Q}{\pi} - \frac{V_Q I}{I_{max}} = 0$$

[illegible]

$$I = I_{max} \frac{2}{\pi}$$

[illegible]

$$P_{Trans}(max) = V_Q\ I_{max} \frac{2}{\pi^2}$$

The most adverse ratio of $P_{Trans}(max)$ to $P_{out}(max)$ is 1 to 2.5. Whereas for class A it is 1 to .5. Thus the real advantage of class B becomes apparent. That is, the heating in the transistors is less than for class A. These relations are summarized in Figure 4.

As a consequence of the conditions described above, we find class A used only at the low to intermediate power levels. Perhaps to a watt or so. And class B used at all higher levels. In fact we usually find class B output stages in linear integrated circuits which operate at only a 100 mw or so.

2. Class A Stages for Low Power Output and Driver Stages

Considerable use is made of class A push-pull stages for [illegible] stages in low voltage applications such as automobile [illegible]

[illegible] class A stage is very easy to design in an approximate [illegible] recognize the following criteria.

1. The output power desired
2. The impedance of the load
3. The supply voltage
4. In lieu of 3, the maximum allowed transistor voltage.

A first design proceeds as follows. A typical D.C. and A.C. load line are shown in Fig. 5. The Q point dissipation must be at least twice the desired output. Thus

$$I_Q V_{CC} \geqq 2 P_{out}$$

The load presented to the transistors should be as large as possible. However there is no point in making V_{max} more than $2V_{CC}$ since we can't use the positive swing anyway. Thus we draw an A.C. load line through the Q point and $2V_{CC} = V_{max}$. The slope of the load line is

$$R_L' = \frac{V_{CC}}{I_Q}$$

The load is R_L. Thus the transformer turns ratio should be

$$n = \left(\frac{R_L'}{R_L}\right)^{1/2}$$

and the design is finished.

Figure 6 shows a typical driver stage. Appropriate bias resistors must be selected and V_E must be allowed for in the overall voltage available. The power gain of the stage is

$$\frac{V_{out} I_{out}}{V_{in} I_{in}} = \frac{I_{out} \; R_L' \; I_{out}}{I_{in} \; R_{in} \; I_{in}}$$

$$= \frac{(I_{out})^2}{(I_{in})^2} \; \frac{R_L'}{h_{ie}} = h_{fe}^{\;2} \; \frac{R_L'}{h_{ie}}$$

This form of circuit is convenient to transformer couple to a higher power push-pull stage since a center tapped secondary can be used to provide the push-pull (phase inverted) drive required.

3. The Class B Push Pull Output Stage

A basic circuit and load line plot for a class B push-pull output stage is shown in Fig. 7. The maximum output is

$$P_{LB}(\max) = \frac{V_{QB}\, I_{\max}}{2}$$

or

$$P_{LB}(\max) = \frac{V_{QB}^{2}}{2R_L'} = I_{\max}^{2}\ \frac{R_L'}{2}$$

Clearly the maximum voltage rating of the transistor must be $2V_{QB} = 2V_{CC}$. Again we make R_L' as large as possible to obtain maximum power transfer. The limit on R_L' is determined entirely by the power supply~~ing~~ voltage which must be less than one-half $V_{C\max}$ and the power demanded.

The output transformer turns ratio is simply

$$(R_L'/R_L)^{1/2} = n$$

where n is 1/2 primary to secondary turns ratio.

At the input we must try to match the input impedance of the stage which is

$$R_{iT} = h_{ie} + h_{fe}\, R_E$$

Usually this is just $R_{iT} \cong h_{ie}$. The input turns ratio is then

$$\left(\frac{R_{1T}}{R_{1A}}\right)^{1/2} = a$$

where a is 1/2 secondary to primary turns ratio. This completes the very basic amplifier design.

The next matter is to establish the Q point and hold it reasonably steady as a function of ambient temperature changes and thermal heating of the transistors. This is most often accomplished by using a diode to bias the circuit as shown in Fig. 8.

The resistor R_B supplies a steady current

$$I_B = \frac{V_{CC}-V_D}{R_B}$$

The diode voltage will match that required by the transistor quite well. That is, for Silicon

$$V_D \cong V_{BE} \cong 0.6 \text{ volts}$$

Also the thermal coefficient of voltage will track quite well. This bias arrangement is usually stable by itself as long as the diode is thermally connected to the heat sink of the transistor.

There must be some small steady state bias current, I_Q, used in order to avoid the rather non-linear first part of the voltage transfer curve as shown in Fig. 9. There is actually a rather small range of suitable bias voltage because at higher bias there is undesirable heating of the devices and at lower bias so called "crossover distortion" occurs.

Waveshape changes with bias are shown in Fig. 10. Usually a substantial amount of feedback is used to reduce the residual distortion in amplifiers of this type.

The gain of the class B stage is

$$\text{Power Gain} = h_{fe}^{2} \frac{R_L'}{R_{iT}}$$

This can easily be several thousand. Thus, a 10 watt power stage can be driven by a 10 milliwatt driver stage. The driver stage will usually be a class A type, single ended.

4. Direct Coupled and Complementary Power Amps.

Because of the size, cost, availability and poor frequency response of transformers, the transformer coupled push-pull form of amplifier has given way largely to the direct coupled transistor types described briefly in this section.

Two of the more common forms of power output stages used are shown in Figs. 11 and 12. In Fig. 11 we see the simplest form of complementary circuit. This makes use of an NPN and PNP power transistor and an NPN driver stage. The details of bias, voltage swing and power capabilities will be discussed in more detail later. In Fig. 12 we see the so called quasi-complementary form. This uses a compound transistor in NPN-NPN form and a lower compound made up of a PNP followed by an NPN. The reasons for compounding are of course to obtain very high current gain. The not so obvious reason for the lower complementary-compound stage is the difficulty of finding suitably high power PNP devices. Thus the PNP-NPN compound has the electrical appearance of a PNP transistor with the high current handling capabilities of the NPN final stage.

Our discussion will be concerned with first a general picture of operation and second with more detailed voltage, current and power limitations.

Let us first consider the complementary output stage shown in Fig. 11a. The output stage is biased with two diodes so the transistor is in class B. That is right on the edge of conduction. The driver is biased via an appropriate feedback network so that

it operates class A at a collector voltage approximately [illegible] between V^+ and V^-. With $V^+ = V^-$ this puts the output at ground for D.C. bias conditions. The input signal is A.C. coupled to the driver which provides both voltage and current amplification. The output stage then drives the load through T_1 on the positive part of the cycle and through T_2 on the negative part. Thus we could provide drive almost to V^+ and V^- on the respective half cycles. In some schemes, we use only one power supply and couple to the load through a capacitor. This simplifies the power supply but requires a large capacitor in series with the load if the load is low resistance.

In this amplifier, both T_1 and T_2 are often compounded to get high current gain. For compound connection we have approximately

$$h_{ie} = h_{ie1} + h_{ie2}\, h_{fe1}$$

$$h_{fe} = h_{fe1}\, h_{fe2}$$

Since the voltage gain of the output stage is unity, we have a power gain of about

$$1 \times h_{fe} = h_{fe1}\, h_{fe2}$$

This can easily be several thousand. Thus the class A driver can operate at a fraction of a watt in an amplifier which delivers 100 watts or more. More exact limits on the gain and bias situation are discussed a bit later. When the output transistors are compounded, 4 diodes are used to provide bias. In addition small emitter resistors are introduced to control bias and linearize

the transfer curves of the transistors even more.

The quasi-complementary amplifier operates in a similar manner. We will concentrate our attention on the circuit of Fig. 13. For the sake of simplicity we have used a single supply and capacitor coupled to the load. Some values for a fairly high power amplifier and compounded transistors are given.

The expected output waveform at the emitters is shown.

From the figure we write

$$V_{OP}^{+} = V_{CC} - V_Q - 2V_{BE} - I_{BP}^{+} (R_x + h_{ie}) \qquad (1)$$

and

$$V_{OP}^{-} = V_Q - V_E - V_{DSAT} - 2V_{BE} - I_{BP}^{-} h_{ie} \qquad (2)$$

If we make $V_{OP}^{+} = V_{OP}^{-}$ we have

$$I_{out}^{-} = \frac{V_{OP}^{-}}{R_L} = I_{out}^{-} = \frac{V_{OP}^{+}}{R_L}$$

and

$$I_{BP}^{-} = \frac{I_{out}^{-}}{h_{fe2}} , \quad I_{BP}^{+} = \frac{I_{out}^{+}}{h_{fe1}}$$

or for $h_{fe1} = h_{fe2} = h_{fe}$ we have

$$I_{BP} = \frac{V_{OP}}{R_L h_{ie}}$$

Now if we set V_{DSAT} = small $\cong 0$ and sum all the voltage applying with the symmetry condition on $V_{OP}^{+} = V_{OP}^{-}$. We can write

$$V_{OP}^{+} + V_{OP}^{-} + V_E + 2V_{BE} + I_{BP}^{-} h_{ie} + I_{BP}^{+} (R_x + h_{ie}) + 2V_{BE} = V_{CC}$$

and

$$V_{CC} = V_{OP} \left(2 + \frac{R_x + 2h_{ie}}{R_L h_{fe}}\right) + V_E + 4V_{BE}$$

or solving for V_{OP} we find

$$V_{OP} = \frac{V_{CC} - V_E - 4V_{BE}}{2 + \dfrac{R_x + 2h_{ie}}{R_L h_{fe}}}$$

This is one of our design equations. We also can write

$$V_Q = V_E + 2V_{BE} + V_{OP} \left(1 + \frac{h_{ie}}{R_L h_{fe}} \right) \qquad (2)$$

Same as above

or

$$V_Q = V_{CC} - 2V_{BE} - V_{OP} \left(1 + \frac{R_x + h_{ie}}{R_L h_{fe}}\right) \qquad (1)$$

Clearly the last two equations are redundant and may serve as a check on the arithmetic.

If we inspect these equations we find that the output swing, and consequently the power to the load can be made larger by making R_x smaller. Clearly however, smaller values of R_x increase the dissipation in the driver transistor. Thus we must establish as a major criterion in the design the maximum power to be dissipated in the driver.

Since the driver is class A, this occurs at quiescent conditions. After we establish this power level, we may get an approximate value of R_x from

$$R_x = \frac{(V_{CC} - V_Q - 2V_{BE})(V_Q - V_E - 2V_{BE})}{P_Q}$$

Setting $V_Q = \frac{1}{2} V_{CC}$ is a good first approximation since the variation $dP_Q \int dV_Q$ is small at $V_Q = \frac{1}{2} V_{CC}$.

Then the first trial gives

$$R_x = \frac{\frac{1}{4}V_{CC}^{2} - V_{CC}(\frac{1}{2}V_E + 2V_{BE})}{P_Q}$$

For the example values shown in Fig. 13 we proceed as follows.

$$R_x = \frac{\frac{1}{4}(100)^2 - 100(\frac{1}{2}\,3 + 2 \times .7)}{1} = 2200\ \Omega$$

and

$$V_{OP} = \frac{100 - 3 - 2.8}{2 + \frac{2200 + 2000}{2000 \times 8}} = \frac{94.2}{2.262} = 41.5 \text{ volts}$$

Also

$$V_Q = 3 + 1.4 + 41.5\left(1 + \frac{1000}{8 \times 2000}\right) = 48.8 \text{ volts}$$

Thus the first trial is really very good and we can use the design in this form.

As an exercise it is interesting to see how V_Q and V_{DP} depend upon R_L and P_Q. Try $R_L = 4, 8, 16$ and $P_Q = .5, 1, 2$ watts.

For the above amplifier we have in summary,

$V_{OP} = 41.5$ $\quad$ $I_{OP} = 5.2$ amps

$V_Q = 48.8$ $\quad$ $P_{out} = 108$ watts

$R_x = 2200$

In fact, these figures are a bit idealized and optimistic. Some practical considerations which affect our results and add losses in swing are interesting. Since the power out goes as the square of the swing, each volt becomes important.

A more complete amplifier is shown in Fig. 14.

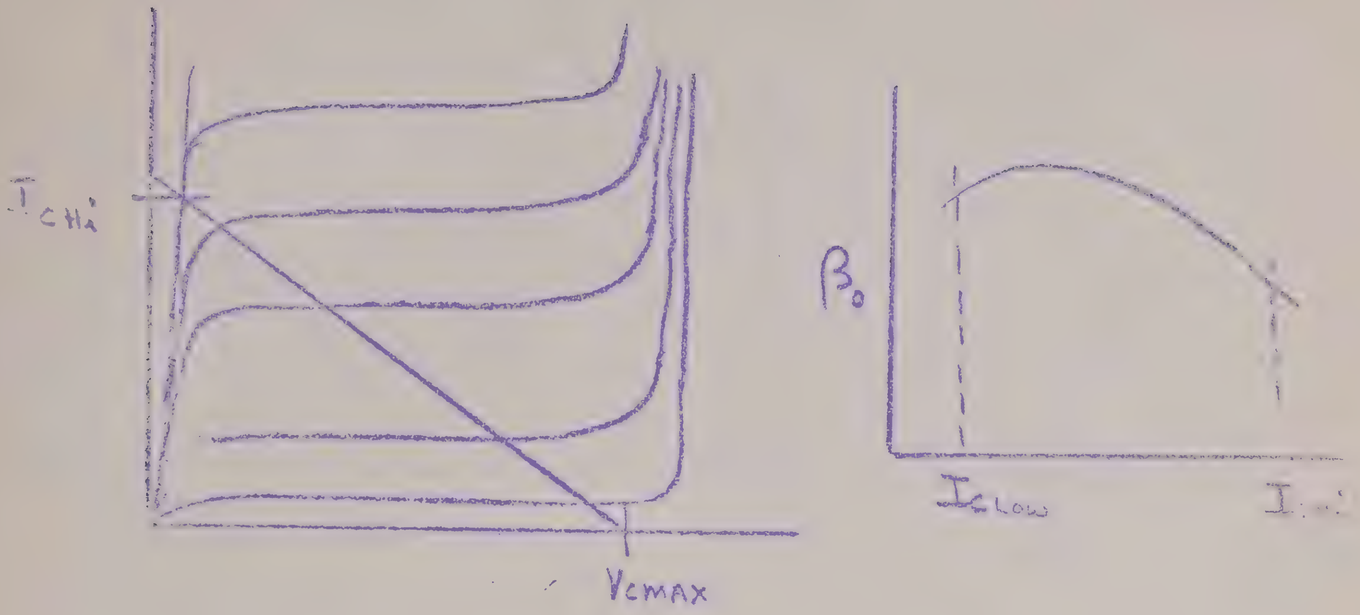
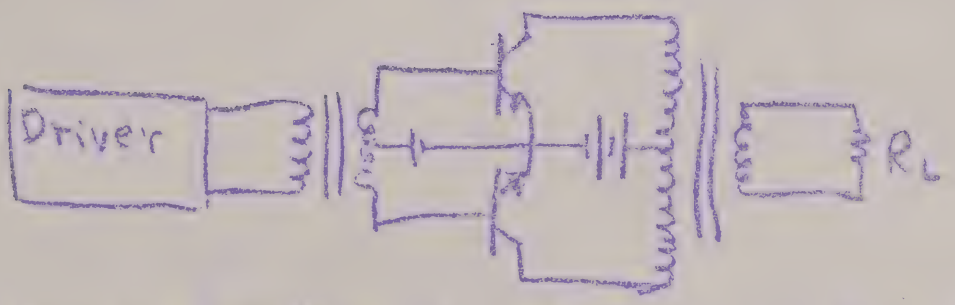
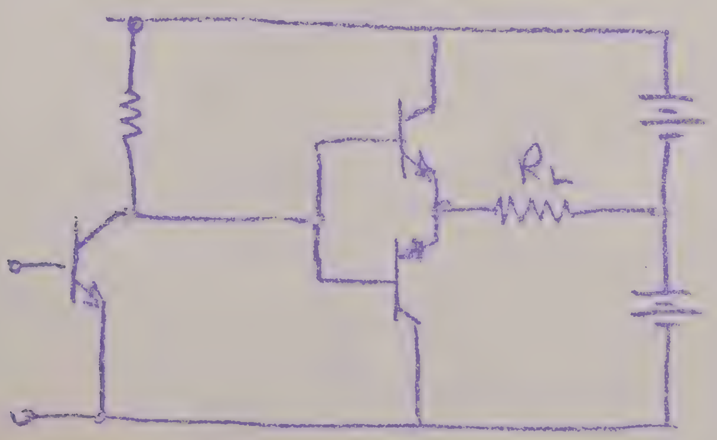


Fig 1.



Common Emitter Push Pull class A or B



Complementary Class B push pull

Fig 2

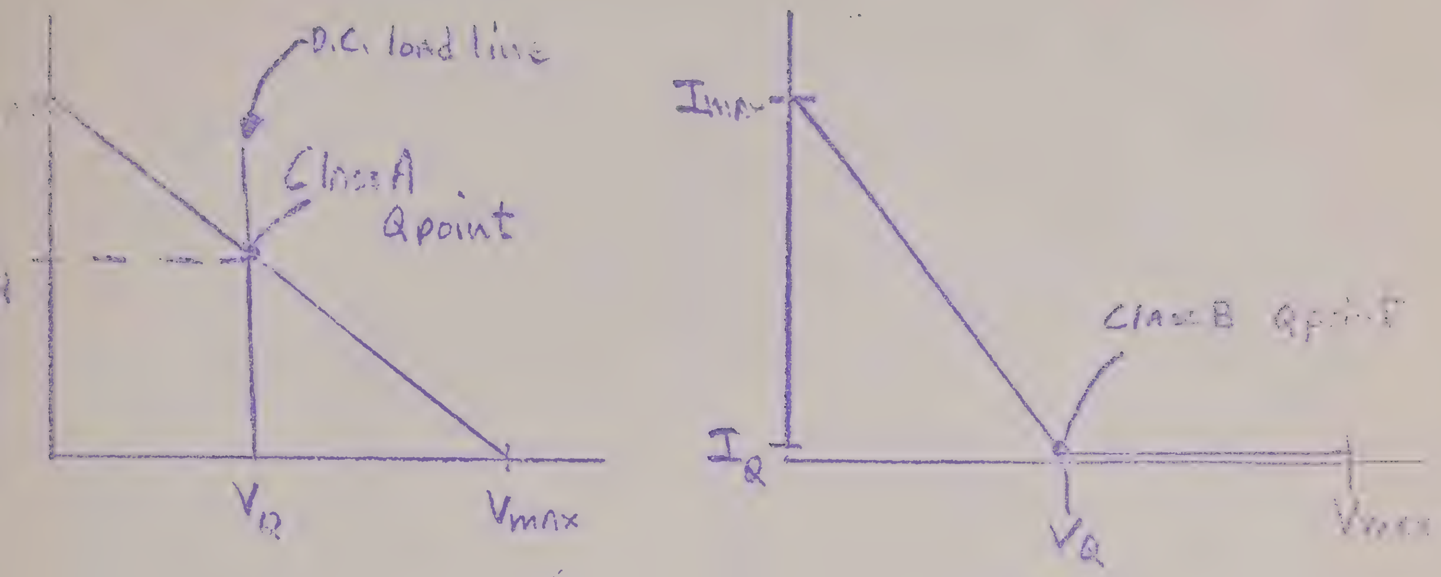

D.C. load line
Class A Q point
V_Q
V_{max}
I_{max}
Class B Q point
I_Q
V_Q
V_{max}

Fig 3.

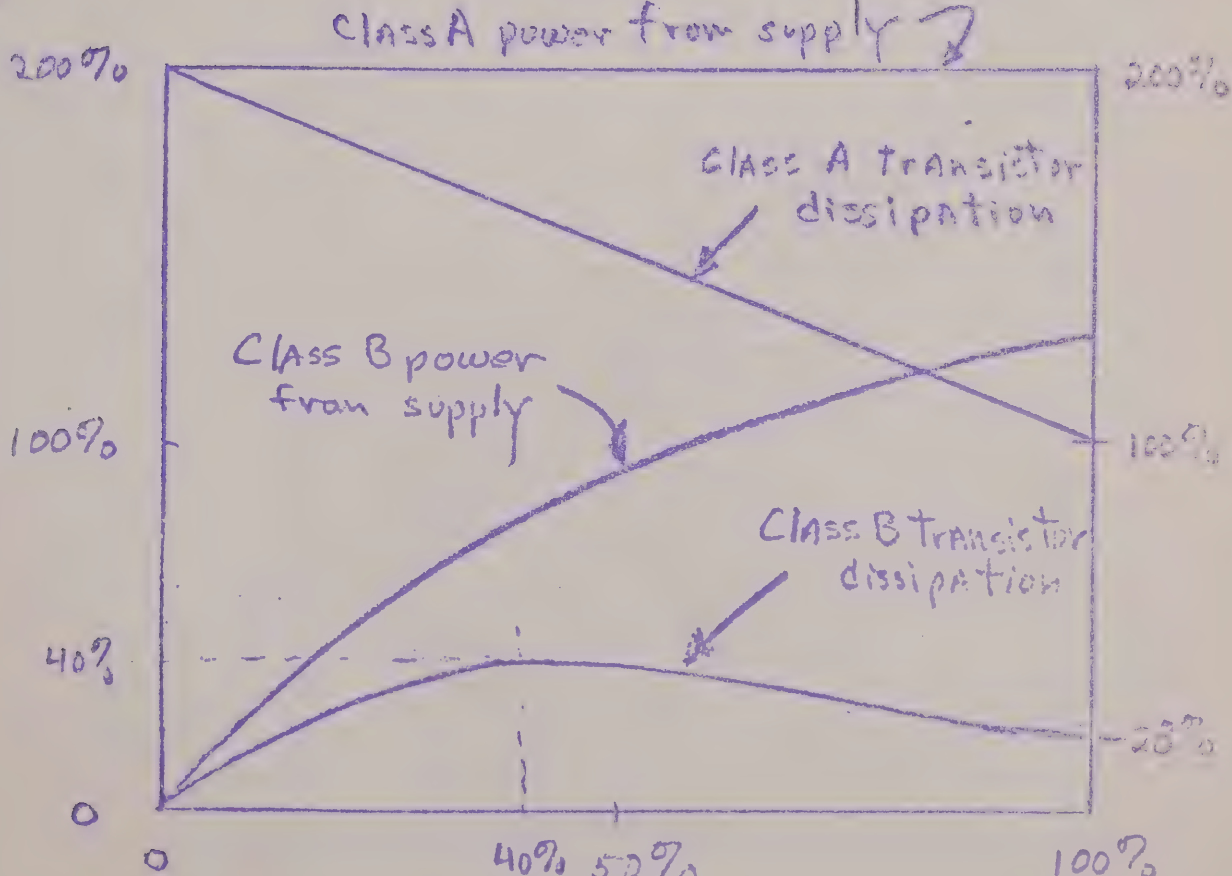

Class A power from supply
200%
200%
Class A transistor dissipation
Class B power from supply
100%
100%
Class B transistor dissipation
40%
20%
0
0
40%
50%
100%

Fig 4

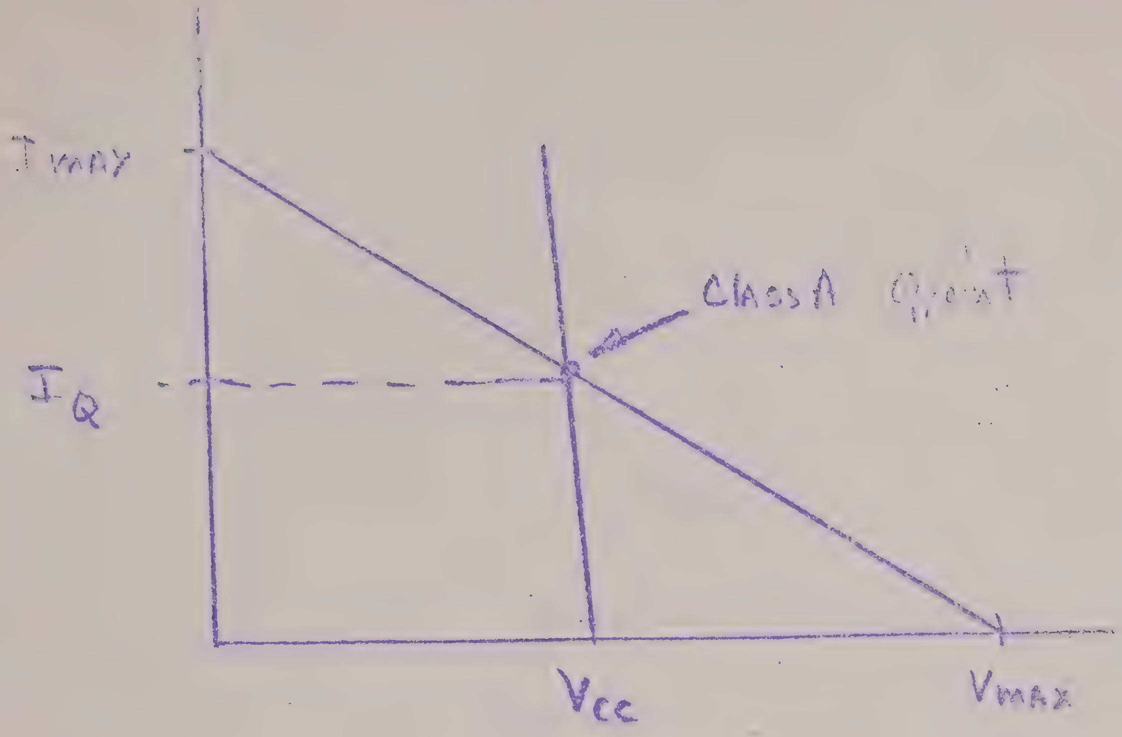
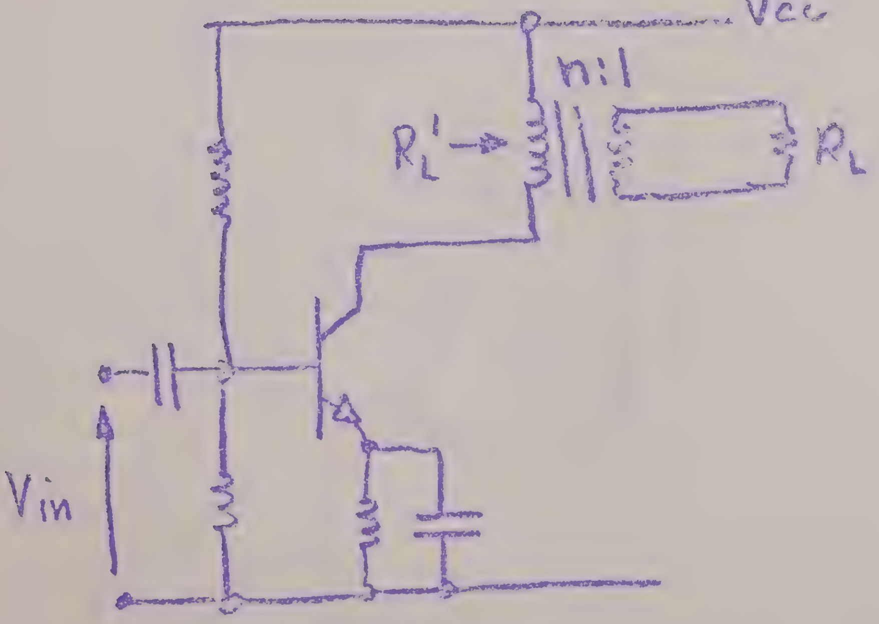
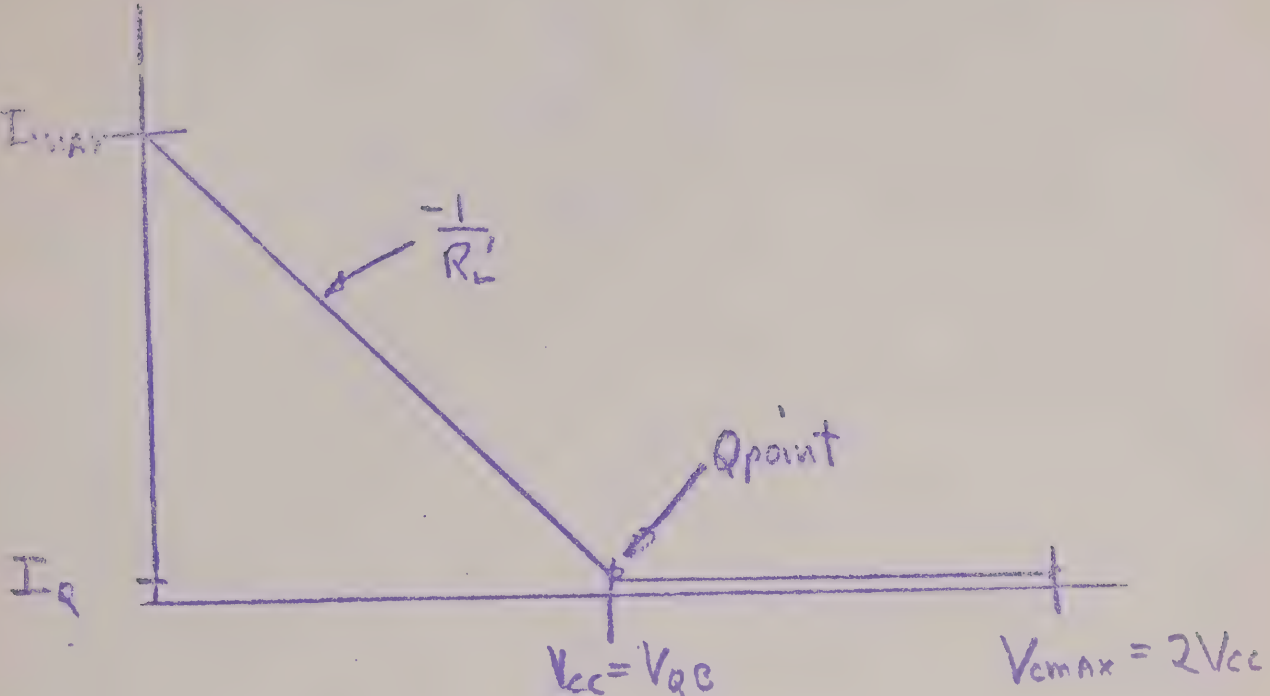


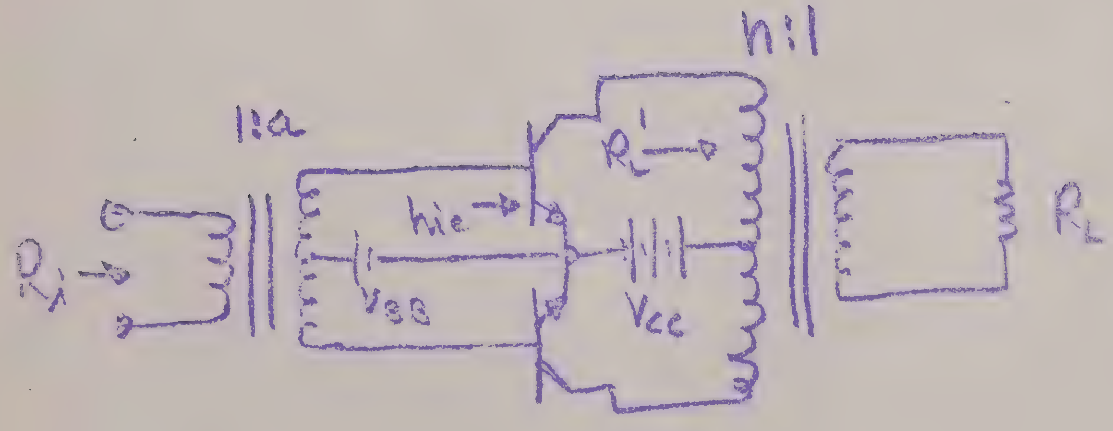
Fig 5,



Class A transformer coupled circuit

Fig 6





Class B push pull amplifier. Transformer coupled and driven.

Fig 7

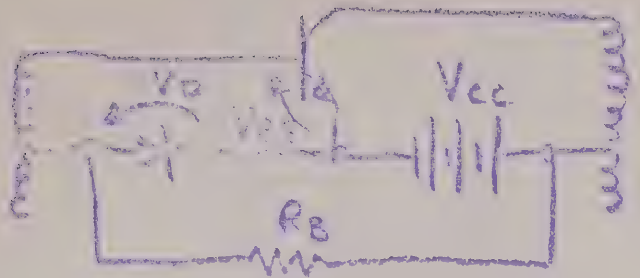


Fig 8

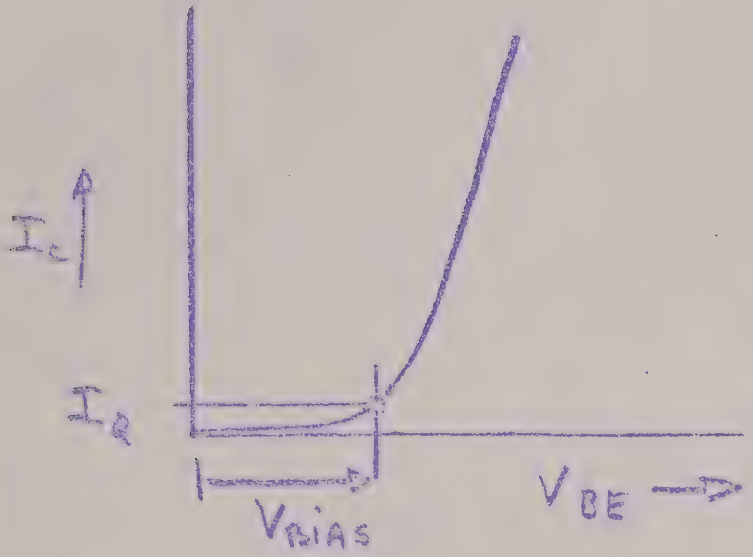
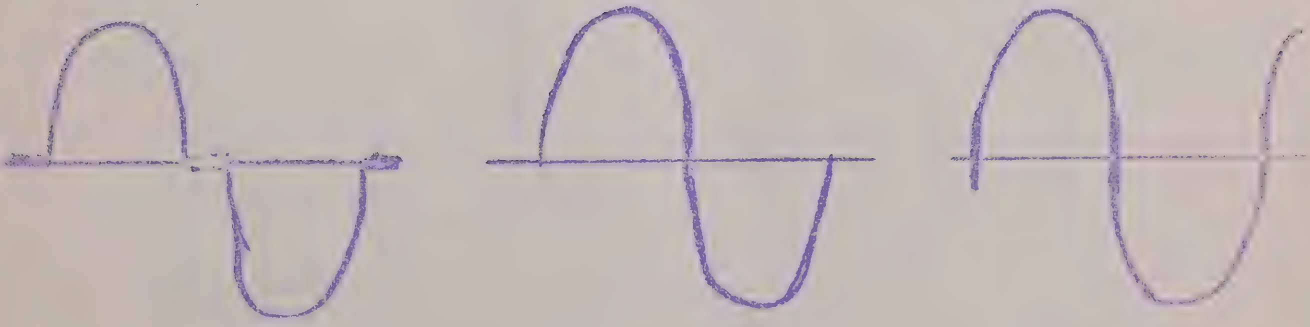


Fig 9



Under Correct Over

Fig 10. Bias conditions

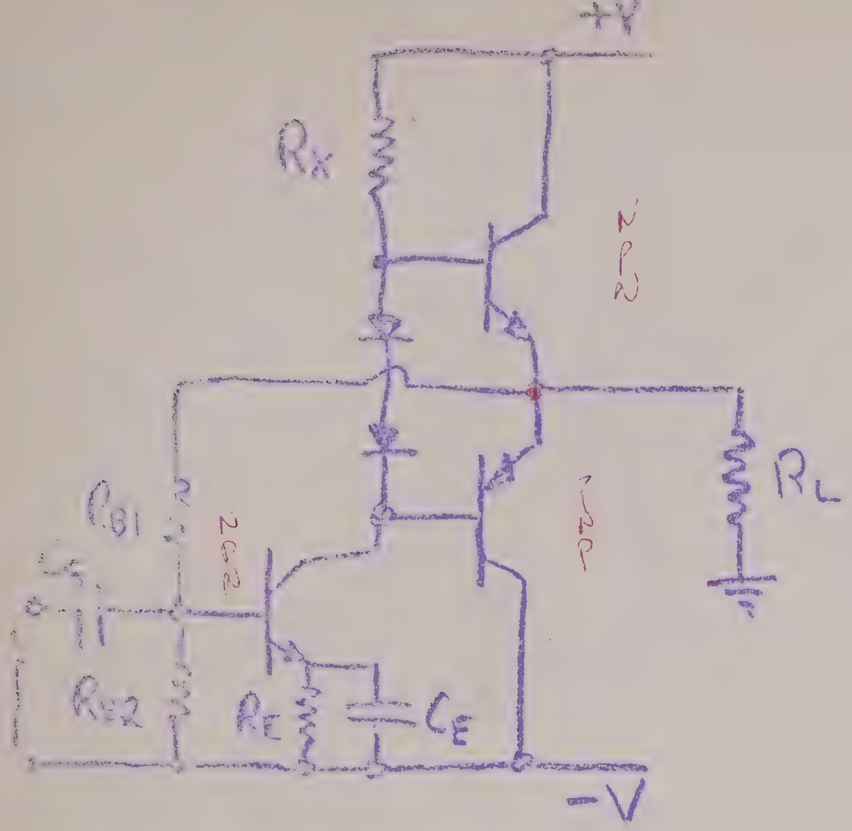


Fig 11

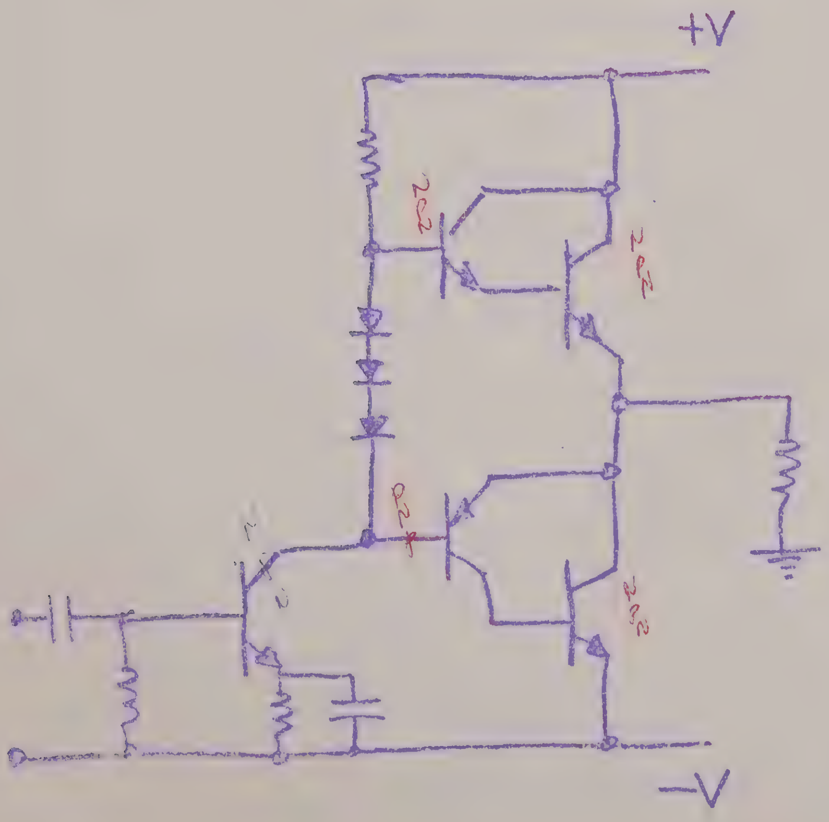
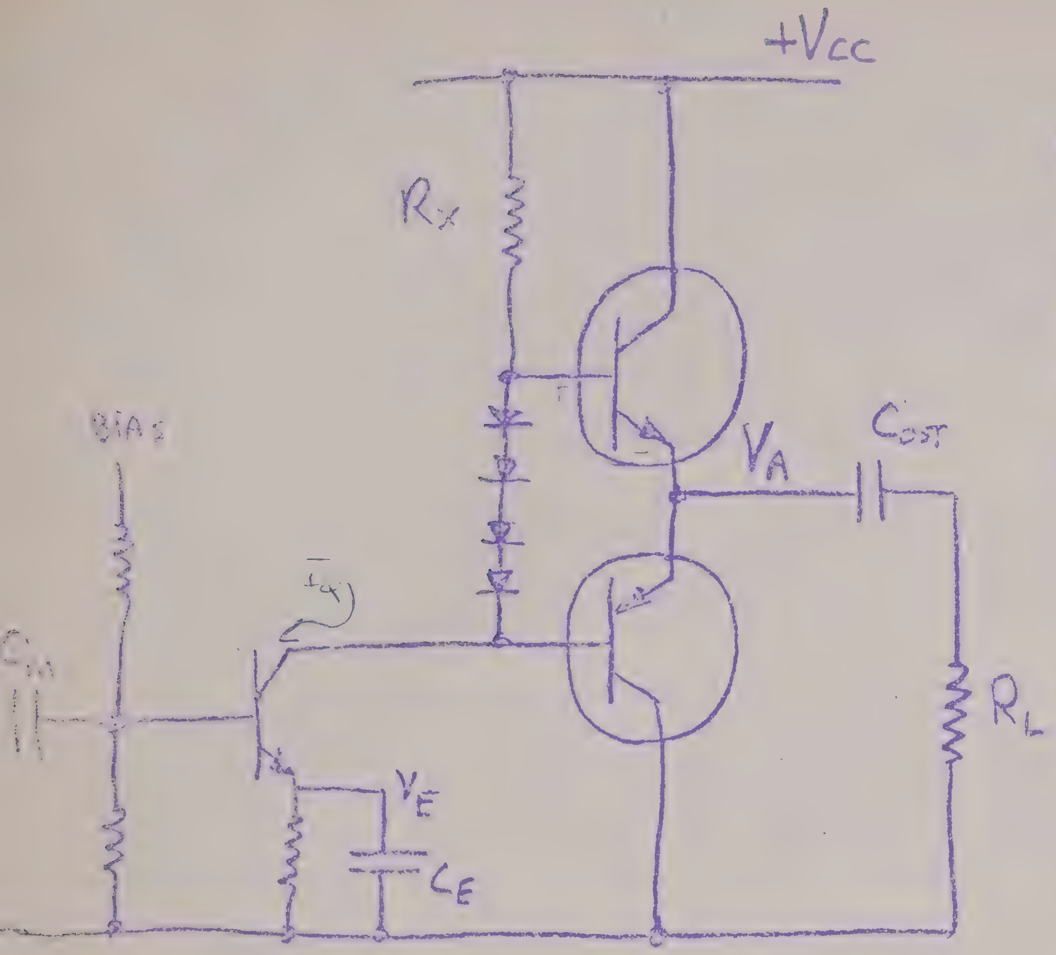


Fig 12



V_{cc} = 100 volts

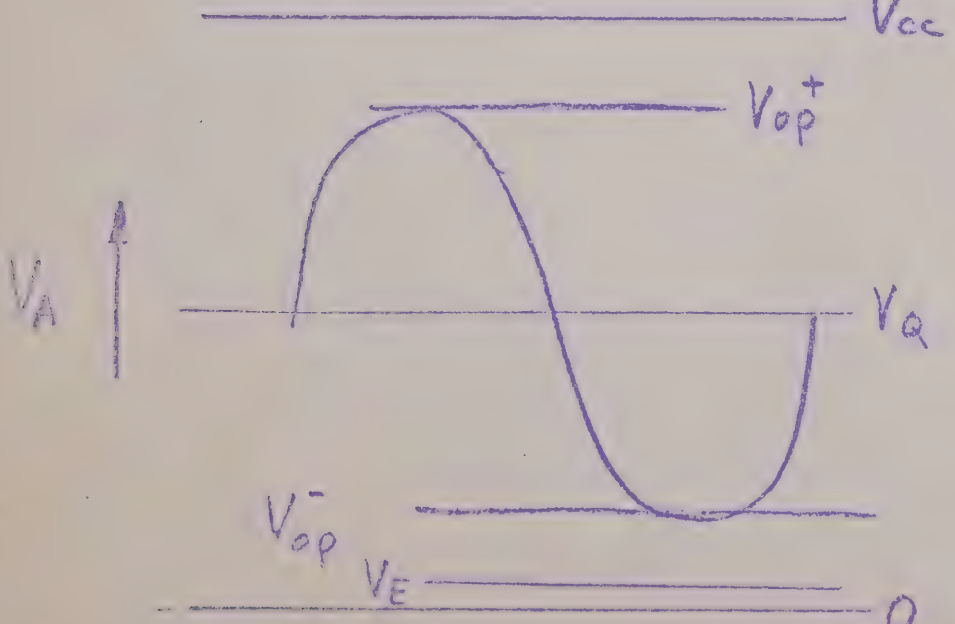
h_{fe} = 2000

h_{ie} = 1000

V_E = 3v

R_L = 8 Ω

P_Q(driver) = 1 wt



Harmonic Oscillators

One of the sources of signals used in electronic apparatus is the oscillator. This is an electronic network which requires only the dc bias as an energy source, i.e., no driving signal is required or needed as in the case of an amplifier. There are two different types of oscillators; the harmonic oscillator whose useful output signal voltage and current are ideally sinusoidal in time, and the switching oscillator whose output voltage and current are periodic but not sinusoidal in time. The discussion here will be limited to the harmonic oscillator.

1. Properties of Harmonic Oscillators

Generally, there are three properties of a harmonic oscillator which are of fundamental importance. These properties are: (a) purity of waveform, (b) stability of generated frequency, and (c) stability of the signal amplitude. Each of these is worth a short discussion at this time.

A harmonic oscillator output is usually used in either an instrumentation application or in a communications system. In each instance, the spectral content of the signals being processed is of first importance. If the oscillator supplies a pure sinusoidal voltage and/or current, then the spectural content contribution from the oscillator is not only known but is as simple in a spectral sense as is possible.

The output of the oscillator is usually used for a time or frequency reference in the above schemes. It is then necessary that the frequency produced by the oscillator be as stable as possible over both short and long time intervals. A good harmonic oscillator may have a frequency stability measured as ±1 part per 10 million, short time stability.

The amplitude of the oscillator output should also be as stable as possible since amplitude changes with time or load causes the equivalent of amplitude modulation which complicates the spectral content of the signal. Thus if there is amplitude instability (jitter), the output waveform will contain signal components in addition to the pure sinusoid desired.

2. Principles of Oscillation

There are several different approaches to the study of harmonic oscillators. The particular one used may be advantageous for some types of oscillator circuits and not too useful for other types, however in principle any one of the approaches may be used. Two of these ways of investigating an oscillator are: (a) feedback theory, Nyquist plots, and infinite gain, and (b) zero impedance-negative resistance. Each approach will be considered.

A. Feedback Theory, Nyquist Plots and Infinite Gain

Consider a basic feedback amplifier such as discussed in Chapter 18, Gray and Searle, pp. 619-620. Equation 18.4 states that

$$V_3/V_1 = \frac{a}{1 + af}$$

For af = -1, the ratio V_3/V_1 tends to infinity. Obviously for V_1 finite, the system will overdrive and the model of the amplifier will not be correct. However, if V_3 is finite, then for af = -1, V_1 must be zero which simply says that this amplifier has an output voltage V_3 without the need of an input voltage V_1. This is exactly what is needed for proper oscillator operation. In order that V_3 represent only a sinusoidal voltage, af must be a function of frequency so that at only a single frequency will af = -1. It then appears that at least the f network should be frequency selective so that af = -1 for the narrowest possible frequency range for best waveform generation.

One may also consider the Nyquist criterion. The Nyquist criterion, usually used for investigating feedback amplifiers and systems for instability, simply states that a simple feedback amplifier will be stable if the magnitude of af(jω) is less than unity when the angle of af(jω) is ±180° (page 729, Gray and Searle). But suppose that af(jω) = 1 when the angle of af(jω) = ±180°; then as an amplifier, the system is unstable, but as an oscillator, the condition for oscillation is met and further, the frequency at which this occurs from the Nyquist criterion will be the frequency of oscillation. For Nyquist consideration, the gain margin = 1.0, and the phase margin is zero at the frequency of oscillation.

B. Zero Impedance and Negative Resistance

Another approach to an oscillator study arises from the knowledge that a network which supplies a signal voltage or current to a load, either real or imaginary, even though the network is supplied from no external signal source, must have exactly zero damping. Thus each physical resistance must be either in series with a resistance of exactly its own magnitude but negative or is shunted by its negative counterpart. Therefore, either a calculation to determine the negative resistance in series with a physical resistor or a net resistance of zero will yield the conditions for oscillations which are the same as those obtained from Nyquist conditions, feedback theory, or infinite gain.

Consider the circuit shown. This is one of the common oscillator configurations. Consider the base resistance $r_x = 0$. Then

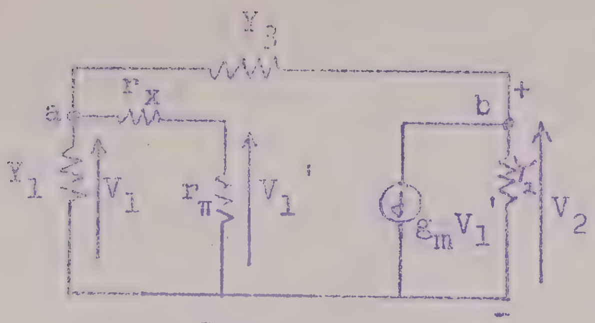


Fig. 1

$V_1 = V_1'$. The nodal equations for this circuit are

$$0 = V_1(Y_1 + \frac{1}{r_\pi} + Y_3) - Y_3V_2 \qquad (1)$$

$$0 = V_1(g_m - Y_3) + V_2(Y_3 + Y_2) \qquad (2)$$

Letting $\frac{1}{r_\pi} = Y_i$, the transconductance is

$$g_m = -\frac{(Y_1+Y_i)(Y_2+Y_3) + Y_3(Y_2)}{Y_3} \qquad (3)$$

Now although g_m is a pure real number, the right side of this equation is in general a complex number. Thus,

$$g_m = G(\omega) + jB(\omega) = G(\omega) + j0 \qquad (4)$$

In general, only one value of ω called ω_o will satisfy Eq. (4) and this equation is used to determine ω_o. Thus, if

$$g_m = G(\omega_o) \qquad (5)$$

the circuit will oscillate since the gain margin will be

$$\frac{g_m}{G(\omega_o)} = 1 \qquad (6)$$

and the phase margin will be zero since

$$jB(\omega_o) = j0$$

Again, consider Fig. 1. It is desired that the impedance connected to Y_3 be determined. With Y_3 removed, the impedance between terminals ab is

$$Z_{ab} = \frac{1}{Y_1+Y_i} + \frac{1}{Y_2} + \frac{g_m}{(Y_1+Y_i)Y_2} \tag{7}$$

If

$$\frac{1}{Y_3} + Z_{ab} = 0 \tag{8}$$

at some value of ω, oscillations will occur. Substituting Eq. (7) into Eq. (8) yields

$$g_m = - \frac{Y_1+Y_i)(Y_2+Y_3)+Y_2Y_3}{Y_3} \tag{9}$$

which is the same as Eq. (3). Note that the net impedance around any loop would have been set equal to zero to determine the condition of Eq. (9).

To give some physical insight into this procedure, consider a specific example. Assume $Y_i = 0$, g_m is a pure real number, and Y_1 and Y_2 to be both capacitive admittances. Thus,

$$Y_1 = J\omega C_1. \qquad Y_2 = J\omega C_2$$

Substitution in Eq. (7) yields

$$Z_{ab} = - \frac{g_m}{\omega^2 C_1 C_2} + \frac{1}{J\omega C_1} + \frac{1}{J\omega C_2} \tag{10}$$

This represents a series connection of two series capacitors and a negative resistance. Eq. (8) states that

$$Z_{ab} + \frac{1}{Y_3} = 0 \tag{11}$$

so

$$\frac{1}{Y_3} = R + J\omega L \tag{12}$$

and then

$$R = \frac{g_m}{\omega_o^2 C_1 C_2} \qquad (13)$$

and

$$\omega_o = \frac{1}{\sqrt{LC_1C_2/(C_1+C_2)}} \qquad (14)$$

Thus, the net damping of the complete circuit is zero at the frequency ω_o even though the resistance R does provide positive damping to the system.

3. Typical Oscillator Circuit

There are many possible oscillator circuits in which the conditions for oscillations can be determined. Only two of these will be considered here.

A. The Colpitts Oscillator

A schematic diagram of a Colpitts oscillator is shown in Fig. 2. For this circuit, the two capacitors designated C are arbitrarily large so that there is no voltage drop across them. The capacitor C in series with R is used simply to keep R from short-circuiting the collector to ground at dc. The resulting equivalent circuit is

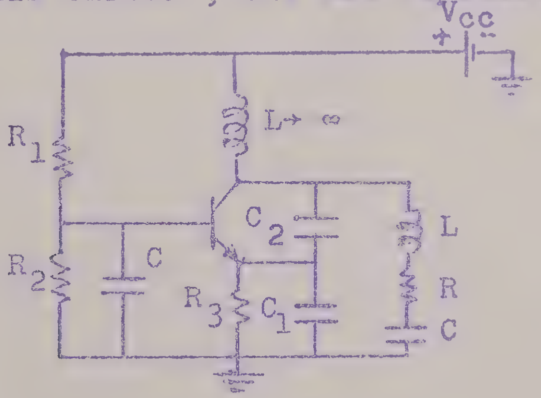


Fig. 2

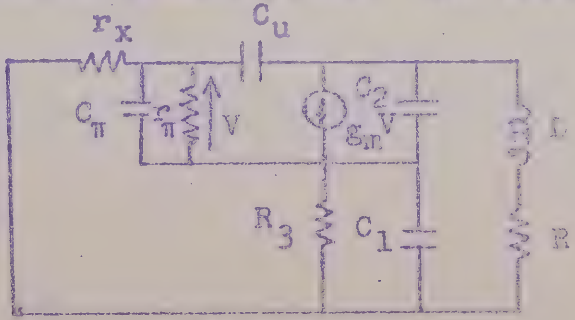


Fig. 3

If the dc collector bias current is not too high and C_1 is large, r_x, r_π, and R_3 can be ignored. The resultant equivalent circuit then is , Fig. 4, and note that C_1 is in parallel with C_π. One of the

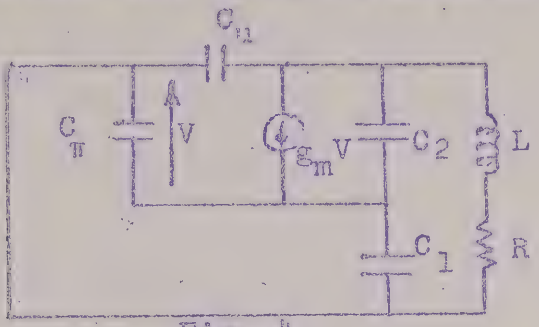


Fig. 4

good properties of the Colpitts oscillator is that C_1 does parallel C_π and can be chosen to so that $C_1 \gg C_\pi$. Similarly, C_2 is also usually made very large to mask any stray capacitance variations between collector and emitter. With C_2 and C_1 of the same order of magnitude and rather large, L must be small and C_μ which shunts the inductor will have negligible effects. Thus, the simplest equivalent circuit, Fig. 5, will be adequate for analysis if C_1 and C_2 are large. This is the same circuit as that of Fig. 1 with the

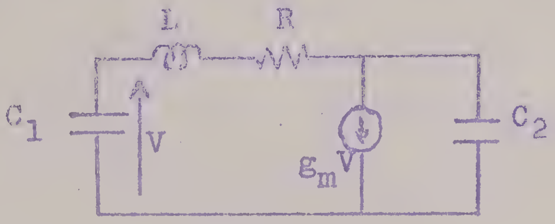


Fig. 5

substitutions used in the specific example. Using those results, Eq. 13 and Eq. 14, oscillations will occur for

$$g_m = \omega_o^2 C_1 C_2 R \tag{13}$$

and

$$\omega_o = \frac{1}{\sqrt{LC_1C_2/(C_1+C_2)}} \tag{14}$$

Consider the following numerical example. Assume that

$C_1 = C_2 = 10{,}000$ pf and $\omega_o = 10^7$ rad/sec. A calculation shows that $L = 2\mu h$, a not unreasonable inductance for $\omega_o = 10^7$ rad/sec. Assume further that for the inductor, $Q_o = \omega_o L/R = 10$, so $R = 2\Omega$. Substitution into Eq. 13 shows that $g_m = 0.02$mho, an entirely reasonable value. However, if $g_m > 0.02$ mho, the system will have negative damping and the signal levels will increase until nonlinear effects limit this signal and the linear equivalent circuit used will no longer apply, but at least the circuit will oscillate.

B. An RC, Phase Shift Oscillator

The Colpitts oscillator previously discussed finds its greatest use in the range of frequencies of 100 KHz to 20 MHz due to limitations on the size of the inductor L. At low frequencies such as audio, RC oscillators are commonly used. One of these, a phase shift type, is shown in schematic form in Fig. 6. Here an FET is used.

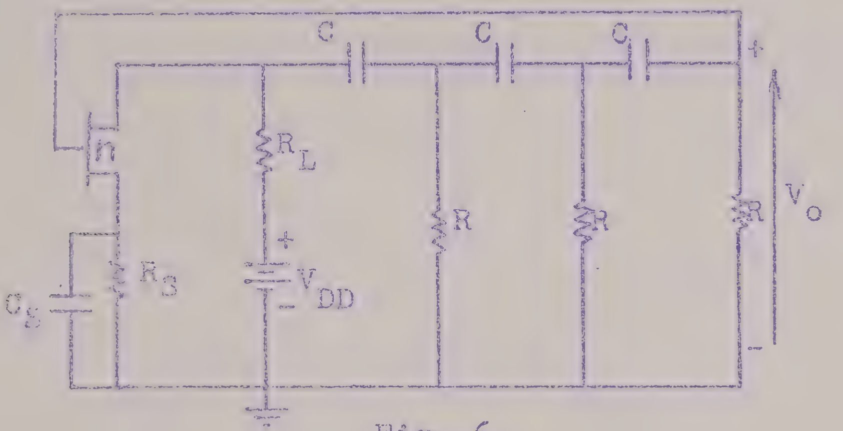


Fig. 6

With this oscillator operating at low frequencies, the equivalent circuit of the FET will contain no capacitances and the overall equivalent circuit is as in Fig. 7. Capacitor C_s is to be large

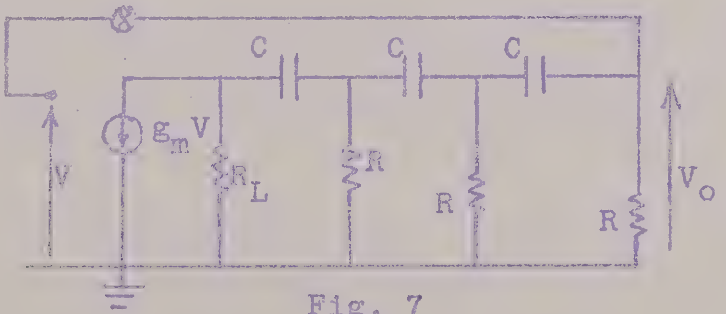


Fig. 7

enough to make the source to ground voltage zero. To analyze this circuit, assume that $R_L \ll R$. Then the open loop gain (G) [illegible]

is

$$\frac{V_o}{V} = \frac{-g_m R_L'}{1 - 5\left(\frac{\omega_1}{\omega}\right)^2 - j\left[6\,\frac{\omega_1}{\omega} - \left(\frac{\omega_1}{\omega}\right)^3\right]} \tag{15}$$

where

$$R_L' = (R_L \| r_d) \tag{16}$$

and

$$\omega_1 = \frac{1}{RC} \tag{17}$$

If the circuit is to oscillate upon closing the loop (ⓧ closed), $V_o/V = 1$. This then requires that the j term in Eq. 15 equal zero. Therefore

$$\frac{6\omega_1}{\omega} - \left(\frac{\omega_1}{\omega}\right)^3 = 0 \tag{18}$$

and solving for ω yields

$$\omega = \omega_0 = \frac{1}{\sqrt{6}} = \frac{1}{\sqrt{6}\,RC} \tag{19}$$

Substituting this value of ω in the remaining part of Eq. 15 gives

$$g_m R_L' = 29 \tag{20}$$

Thus, if $g_m R_L' = 29$, the circuit will oscillate at a frequency of ω_o. If $g_m R_L' > 29$, the circuit will also oscillate but will overdrive itself and again, the linear equivalent circuit will no longer apply.

I. Device Modeling: Real Circuit Components & Linear Circuit Elements

A. Read pages 1 - 10 in the text.

B. An Example - A Flashlight

The typical flashlight consists of an incandescent bulb, one or two dry cells, and a switch, all connected in series. The simplest circuit model for the flashlight is shown in Fig. 1. The dry cells (circuit components) are represented by a voltage source, V_B, (circuit element). The flashlight bulb is represented by a resistance of value R. The source V_B and resistance R are both linear circuit elements. The use of only such elements in a circuit model insures that currents and voltages will be linearly related no matter how complex the model, i.e., no matter how many linear elements it contains and no matter how they are interconnected.

A circuit model must have some purpose. Presumably it should be able to aid us in answering pertinent questions about, in this case, the flashlight. For example we might wish to know:

1. the power dissipated in the bulb,
2. the power dissipated in a bulb which has a different resistance than this one,
3. how much light is produced in both cases,
4. how long the dry cells will last in both cases.

If we have correct values for V_B and R, question number 1 can be easily answered using the model of Fig. 1. For question 2, we can use the model of Fig. 1 only if we use new values for both R and V_B. This is true because the terminal voltage of a dry cell normally decreases as more power is drawn from the cell. Figure 2 shows this for a used "C" cell. Since the volt-ampere relation is roughly linear, an internal battery resistance R_B can be added in series with V_B in the circuit model.

Exercise #1: Using the curve of Fig. 2, draw a new circuit model for the flashlight using correct numerical values for V_B and R_B. Assume that the flashlight uses this one "C" cell. What is the ratio of power dissipated in a 3Ω light bulb to that dissipated in a 9Ω light bulb?

It will be noted that to answer even question 1 correctly with the model of Fig. 1, one cannot use the voltage stamped on the battery case for V_B. The value for V_B must be that measured at the correct current value. The model of Exercise 1 is perhaps more reasonable to use here also.

Questions 3 and 4 cannot be completely answered with any circuit model. The light output of a heated filament depends upon its temperature. This in turn depends upon heat loss by conduction and convection as well as upon the electrical power input. The life of a dry cell depends upon the chemical energy stored as well as the efficiency of conversion to electrical energy. Consequently, either some additional data is needed from the component manufacturer or these questions must be answered by direct experiment.

> Exercise #2: Two "D" cells and a "PR-2" bulb are used in a conventional flashlight. Explain how you could predict a curve of light power output versus time for this flashlight. Indicate what data is needed from the cell manufacturer and from the bulb manufacturer. Is there any advantage to this technique over a direct lifetime measurement?

It should be realized that both R_B and R change with time as the battery discharges. The increase in R_B is fairly obvious. But, the decrease in R is perhaps not so obvious. It is due to the fact that the resistivity of the bulb filament wire is a fairly strong function of temperature.

> Exercise #3: At low temperature the resistance of a certain flashlight bulb is 1Ω. Estimate its resistance in its operating condition. (Look up temperature coefficient of resistance for tungsten and estimate temperature from filament color.)

A fifth question might be asked: "How does the current change with time after the switch is closed?" This is not easy to answer because R changes as the bulb heats. In addition to that, there will be effects due to capacitance and inductance. For example, the filament is probably a coil of wire which will exhibit an inductive effect limiting the rate at which current can build up after closing the switch. An entirely new circuit model is necessary to describe this transient situation.

One last aspect of the flashlight circuit deserves consideration. It should now be evident that the linear model for the bulb has limited value. In reality, if plotted slowly, the volt ampere curve would look like Fig. 3. If, in fact, the curve of Fig. 3 is replotted on Fig. 2 ($V_{bulb} = V_{terminal}$ and $I_{battery} = I_{bulb}$ in the flashlight circuit), the intersection of the two curves gives the answer to question number 1. This is the graphical solution of two simultaneous equations. One is a straight line of the form

$$V = V_B - IR_B \quad \text{[similar to } y = b - xm] \; .$$

The second, $I = f(V)$, has no analytic form.

The purpose of considering the above example is to point out the following facts.

1. A real circuit contains components which usually exhibit a rather complex behavior.
2. Many times certain aspects of this behavior can be approximated by modeling the component with linear circuit elements. This can be very useful because voltage-current relations in linear circuits are derivable with analytical techniques.
3. If the component shows a very nonlinear behavior it can sometimes be handled with graphical techniques.
4. All interesting attributes of a component can rarely be included in a circuit model.
5. The same component can be modeled in a number of ways depending upon how the model is to be used.

C. Electronic Device Modeling - Aims of this Course

In this course we will study the modeling of three important electronic devices: the semiconductor diode, the metal-oxide-semiconductor field effect transistor (MOSFET), and the bipolar transistor. Our emphasis will be on understanding the circuit models in terms of the internal physical mechanisms. Basic device functions will be discussed but detailed circuit applications will be left for later courses. Since all of these devices make use of semiconductor material, we will begin by studying electronic behavior in the bulk semiconductor and then at interfaces between two semiconductors or between a semiconductor and another material.

The aims of this course are to promote:

1. Understanding of the difference between real devices and linear circuit elements and the need for modeling.
2. Ability to create low frequency small and large signal linear circuit models from a volt-ampere characteristic.
3. Understanding of simple graphical circuit solutions for nonlinear components.
4. Understanding of basic electronic device functions including rectification, amplification, switching
5. Familiarity with meaning of general semiconductor terms such as mobility, diffusion, lifetime, doping, hole, energy gap
6. Physical feel for operation of pn diode, MOSFET, and bipolar transistor.
7. Understanding of at least one equivalent circuit model for each of these devices in terms of physical mechanisms of that device.

[illegible]. Simple Circuit Model for a Flashlight

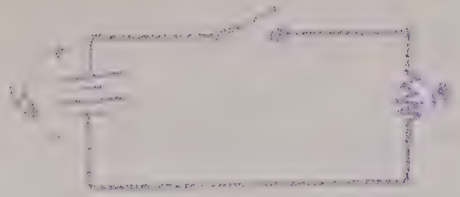


FIG. 2 Terminal Voltage versus Output Current for a Dry Cell

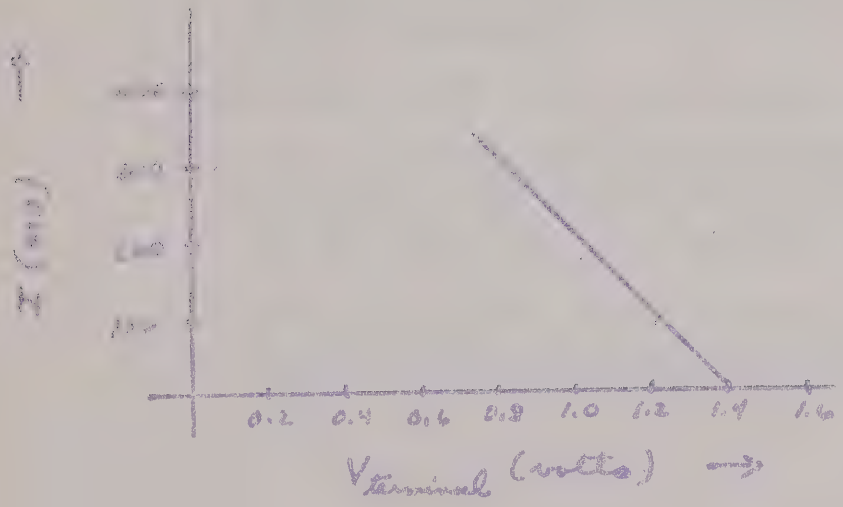
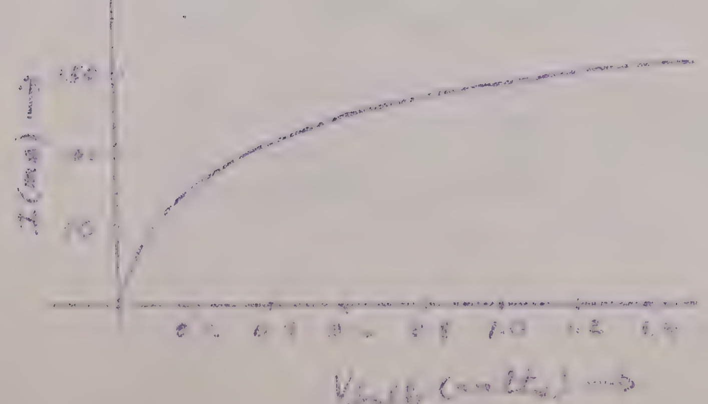


FIG. 3. Volt-Ampere Curve for Light Bulb – Plotted Showi



II. Ohm's Law in Metals and Semiconductors

A. Read pages 43-49. Note the following:

1. In a metal, valence electrons are not bound to any one atom but can move about freely.

2. Typical free elctron density in a metal is of the order of $10^{23}/\text{cm}^3$ and this density is not a function of temperature.

3. The fact that the charge carrier has a negative charge can be proved with a Hall effect measurement.

4. There is no net spece charge in a conductor. The charge of each free electron is compensated by a positively charged ion which cannot move.

5. Any attempt to inject a space charge into a metal results in a rapid redistribution ($<10^{-14}$ sec in duration) such that all excess charge appears only on the surfaces.

B. Read pages 67-70. Note the following:

1. Free electrons in a semiconductor or a metal are in constant random motion at any finite temperature. The average velocity $\overline{v}_e$ in any specified direction is, however, zero in a thermal equilibrium condition.

2. The application of an electric field ε causes a net average drift velocity v_e proportional to ε. The constant of proportionality is called the mobility μ_e.

$$\overline{v}_e = -\mu_e \overline{\varepsilon} \qquad (1) \quad [\text{see Eq. 2.16b}]$$

3. Electrons in metals have smaller mobilities than electrons in semiconductors. (Compare values for Ge, Si, Cu from Problems P2.17, P2.18, P2.19.) This is true because the electrons suffer more frequent collisions in the metal.

4. The semiconductor has an additional positive charge carrier called the hole. It also suffers collisions and has its own mobility.

5. The hole and electron densities in a semiconductor are usually much smaller than the electron density in a metal. In addition, these densities are temperature dependent.

C. Conductivity

As pointed out on page 45, it is possible to write an equation for the current density $\bar{J}$ (current per unit area) due to an electron density n (number of electrons per unit volume) with an average drift velocity $\bar{v}_e$, in the form

$$\bar{J}_e = -q\,n\,\bar{v}_e \,. \qquad (2) \quad \text{[see Eq. 2.1b]}$$

The minus sign is necessary if we consider q to be the magnitude of the charge on an electron. Combining equations (1) and (2) gives

$$\bar{J}_e = q\,n\,\mu_e\,\bar{\varepsilon} = \sigma_e\bar{\varepsilon} \qquad (3) \quad \text{[see Eq. 2.17b]}$$

Equation (3) is Ohm's law in point form for a metal. If it is true for a given resistor, the volt-ampere curve for that resistor is linear. For a semiconductor the equation is expanded to include holes. See Eq. 2.18.

Exercise #4: Using the fact that electric field intensity is defined as the force on a unit charge, show that the power dissipated per unit volume in a conductor is

$$P = \sigma\varepsilon^2$$

D. Resistive Components

Many circuit components exhibit primarily a resistive effect in the circuit. This may be the intended function of the device. One can purchase a large variety of resistors made with metal wire (wire wound), metal thin films and thick films, semiconductor materials (pressed carbon, carbon film, germanium), and small metal grains (cermets). These may be fixed resistors or variable with some type of sliding contact. Other components in this category are resistors used to sense behavior outside the circuit such as temperature sensitive resistors (thermistors, semiconductor resistance thermometers), magnetic field sensitive resistors (magnetoresistors), light sensitive resistors (photoconductors) and pressure or strain sensitive resistors (strain gauges, piezoresistors). Other devices such as the incandescent light bulb, toaster wire, etc., are primarily resistors but their purpose is to provide a useful function outside of the circuit, the circuit merely being used to provide power and the device being a converter of that electrical power to some other useful form--usually heat, light or maybe mechanical power.

E. Forces on Electrons in a Conductor

To understand the behavior of resistive devices one must have some feel for the origins and limitations of Ohm's law. It is well known that if one picks a normal well behaved metal such as copper and forms a resistor that the VI characteristic is linear--but why? The current in the wire is due to electrons moving under the influence of the applied voltage which produces a force on an electron given by

$$\text{Electric Field Intensity} \equiv \text{F per unit charge} = \frac{V}{\ell} \tag{4}$$

where ℓ is the length of the wire. This assumes a uniform wire.

Now if these were simply free noninteracting electrons one might expect that they would accelerate under the influence of the force, i.e.,

$$\overline{F} = m\,\frac{d\overline{v}}{dt} = -q\overline{\varepsilon}\ , \tag{5}$$

where m is electron mass, q is electron charge, and v is electron velocity. By combining Equations (2) and (5), one obtains

$$\overline{\varepsilon} = \frac{m}{q^2 n}\,\frac{d\overline{J}}{dt}\ .$$

If the material is homogeneous with a cross sectional area A, we can write

$$V = \frac{m\ell}{q^2 nA}\,\frac{dI}{dt}\ .$$

This is obviously not Ohm's law. In fact the form of the volt-ampere relation is the same as that of an inductor (although the physical mechanism is different than the inductive effect usually considered in a coil).

Consequently we must conclude that this simple model for electron motion in a metal is wrong. It is wrong primarily because it neglects the fact that an electron cannot pass freely through a material without interacting with it. This interaction can be thought of as a collision process which causes kinetic energy, picked up by the electron accelerating in the electric field, to be transferred to the lattice atoms, causing the material to heat.

The process of transferring energy from the electron cloud to the atomic lattice can be adequately described only with the use of quantum mechanical concepts. One speaks of electron-phonon collisions where the phonon is defined as a quantum of lattice vibrational energy. Such collisions either create or destroy phonons (add or subtract fixed amounts of energy to the vibrational modes). This is analogous to the absorption or emission of light energy from a material where one speaks of electron-photon collisions. In that case the photon is a quantum of electromagnetic energy.

Anyway, if we average the effect over the cloud of electrons, the electron-phonon collisions produce a damping or "frictional" force component which is proportional to average velocity. Consequently, in the presence of a constant electric field the force due to that field and the frictional force become equal in a steady state situation and Equation (1) is justified.

F. Temperature Dependence of σ.

The value of the mobility will depend on temperature. It normally decreases as temperature increases because more phonons exist at higher temperature and therefore electron-phonon collisions become more frequent. This is the cause of the positive temperature coefficient for most metals (and for "extrinsic" semiconductors).

The thermistor exhibits a large negative temperature coefficient of resistance. This is due to the fact that the electron (or hole) density increases rapidly with temperature and this effect overcompensates for the effect of the change in μ. See Eq. (3). This is characteristic of an intrinsic semiconductor. In such a material there are no free electrons at a temperature of 0°K. The material is essentially an insulator for very low temperatures. All electrons are bound to individual atoms in the lattice. However, in contrast to most insulators, it takes a relatively small amount of energy (of the order of 1 electron-volt) to free the outermost bound electrons. Consequently at higher temperatures a significant fraction receive enough thermal energy to be freed so that they can move about in the material as conduction electrons. The temperature dependence of the electron density is approximately

$$n(t) \propto e^{\frac{-E_g}{2kT}} , \qquad (6)$$

where k is Boltzmann's constant, T is absolute temperature, and E_g is the excitation energy necessary to free one electron. This energy also is called the "gap" energy.

Exercise #5: Find the ratio of free electron density at a temperature of 50°C to that at room temperature, 25°C for a semiconductor with a gap energy of: a) 1 electron-volt, and b) 0.5 electron-volt.

III. Charge Carriers in Semiconductors

A. Some important points from pp. 49-63 and lectures.

1. The hole is a particle with positive charge q (=1.6 ×10^{-19} coulomb) and positive mass m_h. It moves about in the semiconductor in much the same way as a "free" electron. For example, in an applied field we can observe a hole mobility μ_h.

2. Electrons and holes can coexist in a semiconductor each having its own mobility.

3. In an intrinsic semiconductor $p = n \equiv n_i$, which is a strong function of temperature. See Equation 6.

4. In group IV semiconductors (Si, Ge), donor impurities (group V atoms) or acceptor impurities (group III atoms) can be added to produce extra electrons or extra holes respectively.

5. This process is called doping and the material is now an extrinsic semiconductor, either n type ($N_d > N_a$) or p type ($N_a > N_d$). The quantities N_d and N_a are densities of donor and acceptor atoms.

6. In the extrinsic semiconductor $p \neq n$. However it is still true that

$$pn = n_i^2(T) \qquad \text{[See Eq. 2.8]} \qquad (7)$$

7. Consider a p type semiconductor. The hole is now called the majority carrier and $p \approx N_a$. The hole density is practically temperature independent near room temperature but

$$n \approx \frac{n_i^2}{N_a}$$

(the minority carrier density) is extremely temperature dependent, as well as being a much smaller number than p.

8. Charge density in a semiconductor with both donor and acceptor impurities is

$$\rho = q(p + N_d - n - N_a) \; . \qquad \text{[See Eq. 2.6]} \qquad (8)$$

B. Energy Band Model for Electrons and Holes in a Semiconductor

Hopefully, you are familiar with the fact that electrons bound to an isolated atom are situated at certain fixed allowed *energy levels*. Classically we like to think of these electrons as traveling in certain fixed orbits about the nucleus. An electron can be excited from one level to another by some external energy source such as electromagnetic radiation (photon absorption). On the other hand, all energies other than these discrete levels are forbidden to the electron.

If a second atom is placed in the vicinity of the first, the two perturb one another in such a way that there are now twice as many energy levels as before. If many more atoms are brought into the picture to form a solid, the effect is to produce bands of energies which include so many closely spaced energy levels that they can be considered as continuous, *allowed* bands. They are in turn separated by *forbidden* bands. See Figure 4. This *band model* is very useful for describing conduction processes in semiconductors and semiconductor devices. In a semiconductor at 0°K all bands up to and including the *valence* band are filled with electrons. The *conduction* band, separated from the valence band by the gap energy, E_g, is empty. See Figure 5.

It is obvious that an empty band cannot contribute to electrical conduction. It is not so obvious that the same is true for a full band. Basically this is a fact because the applied electric field must give energy to the electron cloud in order for there to be a current. Usually the field is not strong enough to excite an electron out of the full band to a higher band. Consequently, in the full band, any increase of energy by one electron must be compensated by a decrease in energy by another. If one wishes to claim that Newton's law holds for these electrons, it is unfortunately necessary to postulate an *effective mass* for them which is a function of their position in the band. In order that there be no net current in the presence of an applied field, those electrons near the top of the band must have negative effective mass if those at the bottom have a positive mass.

Consider an intrinsic semiconductor. At finite temperatures the conduction band is partially filled with a few electrons of density n. They are situated near the bottom of this band, have positive effective mass and can easily be nudged into slightly higher energy levels by an electric field.

In this case the valence band can also be energized by a field because of the empty spaces available after the n electrons per unit volume are thermally excited to the conduction band. Consideration of all of these electrons with varying positive and negative effective masses is however very complicated. Instead of doing this, the superposition argument of Figure 6 is used. One merely equates the almost filled valence band to a completely filled valence band plus another valence band which is nearly empty except for a few particles which have positive charge and positive mass. These particles, called holes, exactly cancel the effects of those negatively charged, negative mass electrons previously added to create a filled band.

Now the filled band can be ignored because it does not contribute to conduction. The valence band becomes an almost empty band of well behaved particles. The density p is equal to n in the intrinsic material. In n type material, the donor atoms contribute extra electrons to the conduction band (and p decreases [page 59]). In p type material the acceptor atoms contribute extra holes to the valence band (and n decreases).

C. Optical Absorption - Excess Carriers

An electromagnetic wave of frequency f contains photons of energy

$$E_p = hf \ , \tag{9}$$

where h is Planck's constant. If E_p is less than the gap energy E_g of a semiconductor, the wave cannot excite electrons from the valence band to the conduction band. If other energy absorbing mechanisms are small, the material will be transparent to that radiation. On the other hand if $E_p > E_g$ the photons can be absorbed. The material becomes opaque for frequencies above $f = E_g/h$.

A pulse of light impinging on a semiconductor can, therefore, produce an excess hole-electron pair density. Obviously these excess holes and electrons will not remain forever but will recombine in a time called the lifetime. If the light is steady, a constant excess density will result, determined by the balance of generation by the photons and recombination.

Exercise #6: Silicon has an energy gap of 1.1 iV. What is the cutoff frequency below which radiation will pass through Si? Name at least two semiconductor materials through which you could probably see.

III. Continued

D. Some points from pp. 63-73.

Most of this material is adequately explained in the text. The most important new relationships presented are Equations 2.10, 2.12, 2.14, 2.19, 2.23 and 2.24. Also remember that, with excess carriers present, $pn \neq n_i^2$.

E. Additional comments on lifetime

Lifetime can be affected by impurities and dislocations in a semiconductor as well as by the basic crystalline structure. For example, highly purified single crystal silicon may have lifetimes in the millisecond (10^{-3} sec) range. Microsecond (10^{-6} sec) times are more common. On the other hand, even very pure GaAs might have nanosecond (10^{-9} sec) lifetimes. In all cases, these times are much longer than the dielectric relaxation time, $\tau_d = \varepsilon/\sigma$.

Let us consider a semiconductor with lifetime τ. If at time $t = 0$ the excess density is $n'(0)$, the decay of excess electrons for $t > 0$ is given by the differential equation:

$$\frac{dn}{dt} + R = 0$$

$$\frac{dn}{dt} + \frac{n'}{\tau} = 0 = \frac{dn'}{dt} + \frac{n'}{\tau} = 0 \quad . \tag{10}$$

This has the solution $n'(t) = n'(0) \exp(-t/\tau)$. This result does not hold if there is carrier transport by drift or diffusion.

IV. The Haynes-Shockley Experiment

A. Introduction

The description of the behavior of excess carriers in a semiconductor usually involves three important phenomena: motion due to an electric field, motion due to a density gradient, and recombination. Since all three phenomena can occur simultaneously and since each can vary with position in the material, the carrier transport problem is, in general, very intricate and [illegible]. It is also mathematically quite complex and consequently, in this course, we choose to use only the results of that mathematics to describe the behavior of a few important devices.

In spite of the mathematical difficulties, a considerable degree of understanding of the physical mechanisms involved in the transport of excess carriers can be obtained through performance of the so-called Haynes-Shockley experiment*. This experiment can be used to measure minority carrier mobility μ, diffusion constant D, and lifetime τ.

B. Injection and collection of excess carriers

The Haynes-Shockley experiment depends upon the ability to excite excess hole-electron pairs at some point or over some small volume of a semiconductor bar. This is normally called injection of minority carriers and it can be effected most easily by shining a spot of light on the bar. Although the majority carrier is also injected in this process, it is rarely mentioned because the injected density is normally much smaller than the thermal equilibrium majority carrier density value. In the future we will show that such minority carrier injections can also be produced by either a pn junction or a certain type of metal-semiconductor junction.

For this experiment it is also necessary to detect the presence, locally, of excess minority carriers in the semiconductor bar. This too can be done with a pn junction or metal-semiconductor (Schottky barrier) contact. In our version of the experiment we place a sharpened point made of phosphor-bronze on the bar. If properly made (witchcraft is necessary here), the contact becomes a rectifier, i.e., it passes majority carrier current easily in one direction but not the other. If the contact is "voltage biased" with the correct polarity (for low conduction), it will exhibit a measurable current increase when minority carriers are present beneath it. In this situation, a high field at the contact will sweep the minority carriers into the detector circuit.

C. The experimental arrangement

The circuit used for this experiment is shown in Fig. 7. The strobotac produces a train of bright light pulses less than 10^{-6} sec in duration. Consequently, at given times, a pulse of holes is produced at $x = 0$ in the bar. A power supply V_o is attached to the ends of the bar with ordinary, non-rectifying ("ohmic") contacts. The detector circuit consists of the point contact at $x = L_1$, a series resistor, and a bias battery. The voltage across the resistor is measured with an oscilloscope.

D. Measurement of μ_h

The power supply produces an electric field in the bar of magnitude

$$E = V_o/L_1 \quad .$$

From the oscilloscope picture it is evident that the pulse takes t_1 seconds to travel from $x = 0$ to $x = L_1$. Therefore the pulse velocity is

$$v = L_1/t_1$$

The mobility of the pulse is

$$\mu = v/E = L_1L_0/t_1V_o \quad .$$

This is also approximately equal to the mobility of the holes μ_h.

E. Measurement of τ

The total number of excess carriers in the pulse will decay in time as

$$P(t) = P(0)\ e^{-t/\tau} \quad .$$

Note that this is not the hole density. If we take oscilloscope pictures for at least two different spacings L_1, it is possible to obtain an approximate measurement for τ. Figure 8 illustrates the effect of varying L_1. To a good approximation $P(t)$ is proportional to the area under the pulse at the time t.

> Exercise #7: If the area under the pulse in a Haynes-Shockley experiment is halved when L_1 is changed from 1.0 cm to 1.5 cm, what is the carrier lifetime? Assume the mobility has already been measured and is 500 cm^2/volt-sec. The value for E is 40 volt/cm.

F. Measurement of D_h

Notice that we have associated lifetime with the pulse area but not with the pulse height. This is due to the fact that diffusion would cause the pulse amplitude to decrease in time even if the lifetime were infinite. The effect of diffusion is to spread the pulse as is shown in Fig. 9. The combined effect of diffusion and recombination is a spreading and a decrease in area as time progresses.

The diffusion coefficient can be calculated if one knows the mathematical expression for the pulse density as a function of time and space. The expression is included here for completeness:

$$p'(x,t) = \frac{P(0)}{\sqrt{4\pi D_h t}}\ \exp\left[-\ \frac{(x-\mu_h E t)^2}{4 D_h t}\right]\ \exp[-t/\tau] \quad .$$

This expression can be derived from the combination of the expressions for drift and diffusion current, coupled with the equation for recombination and the general requirement of continuity of hole flow. The continuity requirement merely states that hole flow into an infinitesimal volume must equal the net recombination rate.

*References:

1. Haynes and Shockley, Physical Review, vol. 75, p. 591, 1949. (This is the original paper.)
2. Gibbons, Semiconductor Electronics, McGraw-Hill, 1966, pp. 120-125. (A good reference book for this course.)

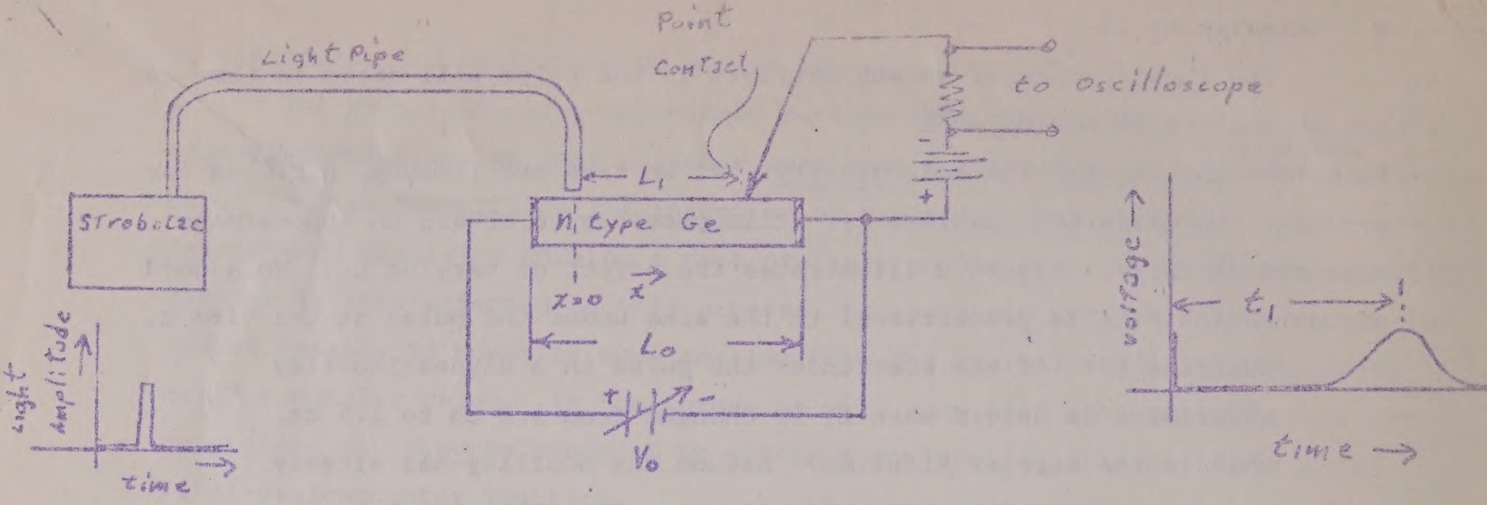


FIG. 7 The Haynes - Shockley Experiment

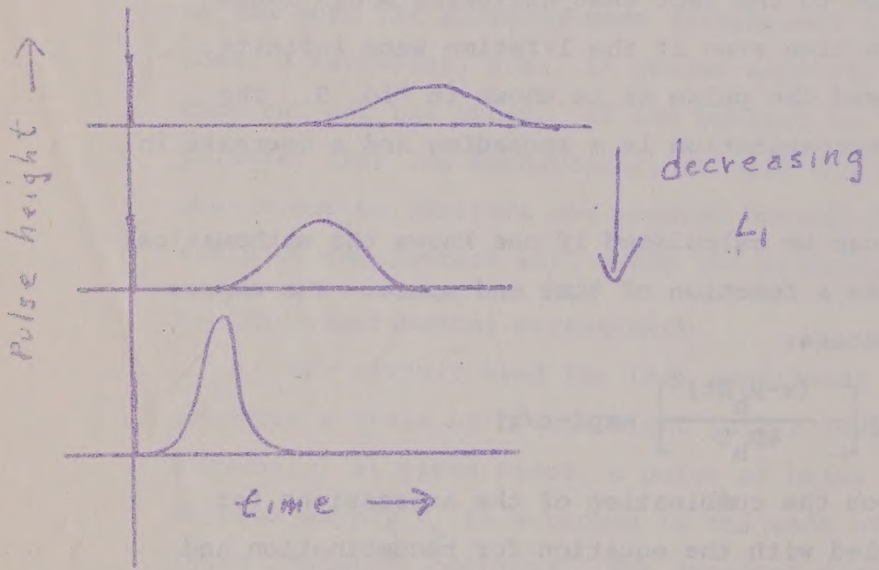


FIG. 8. EFFECT OF A CHANGE IN SPACING L_1 ON SCOPE PICTUR

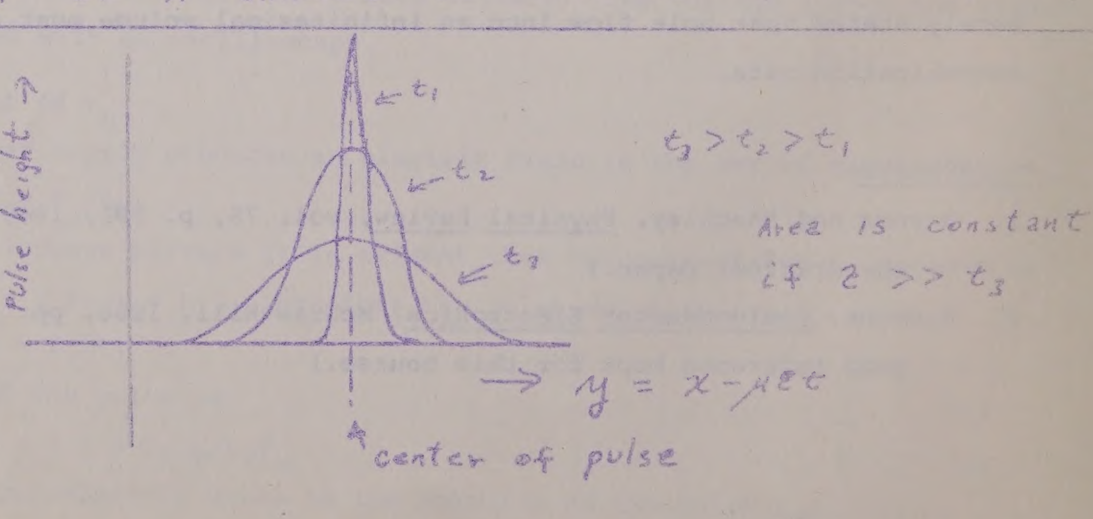


FIG. 9. Effect of Diffusion on Pulse of Excess Carriers

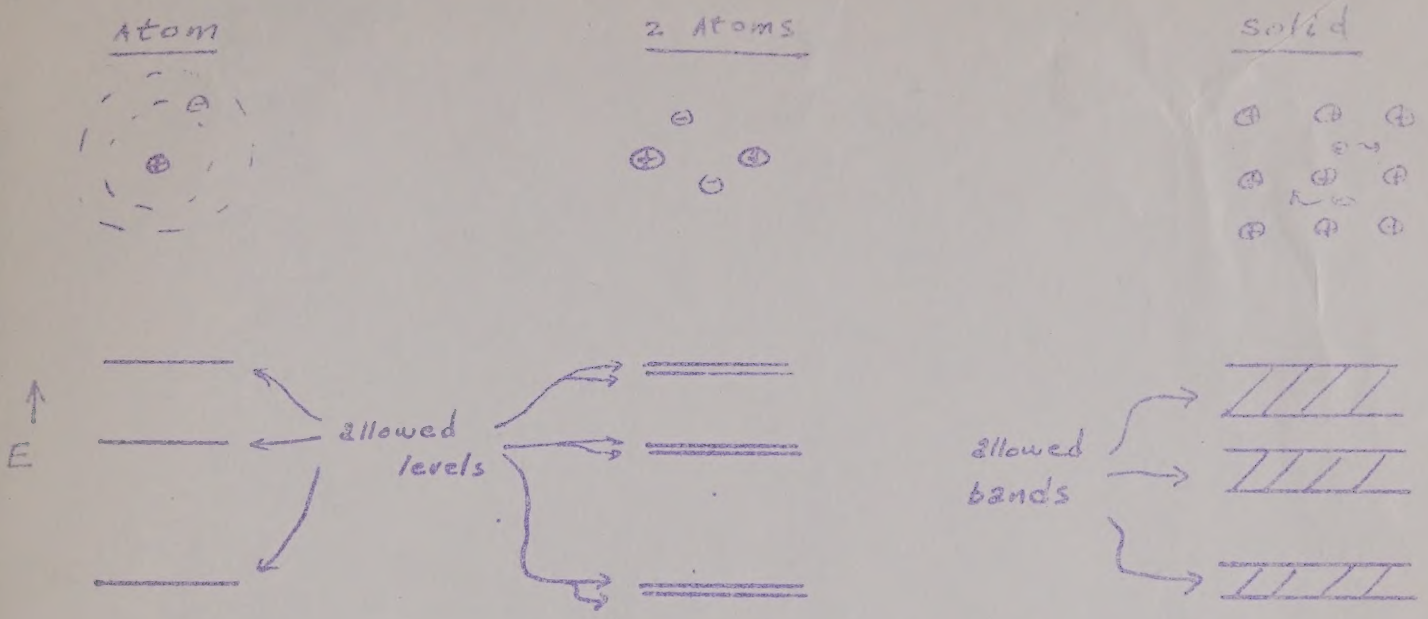


FIG. 4: Electron Energy Levels and Bands

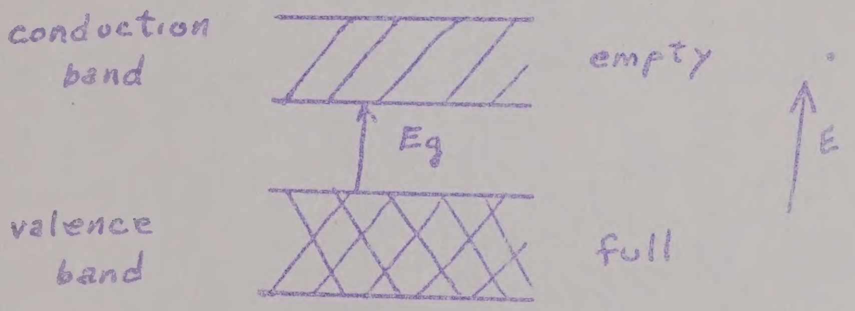


FIG. 5: Electron Energy Band Diagram for Semiconductor at 0°K.

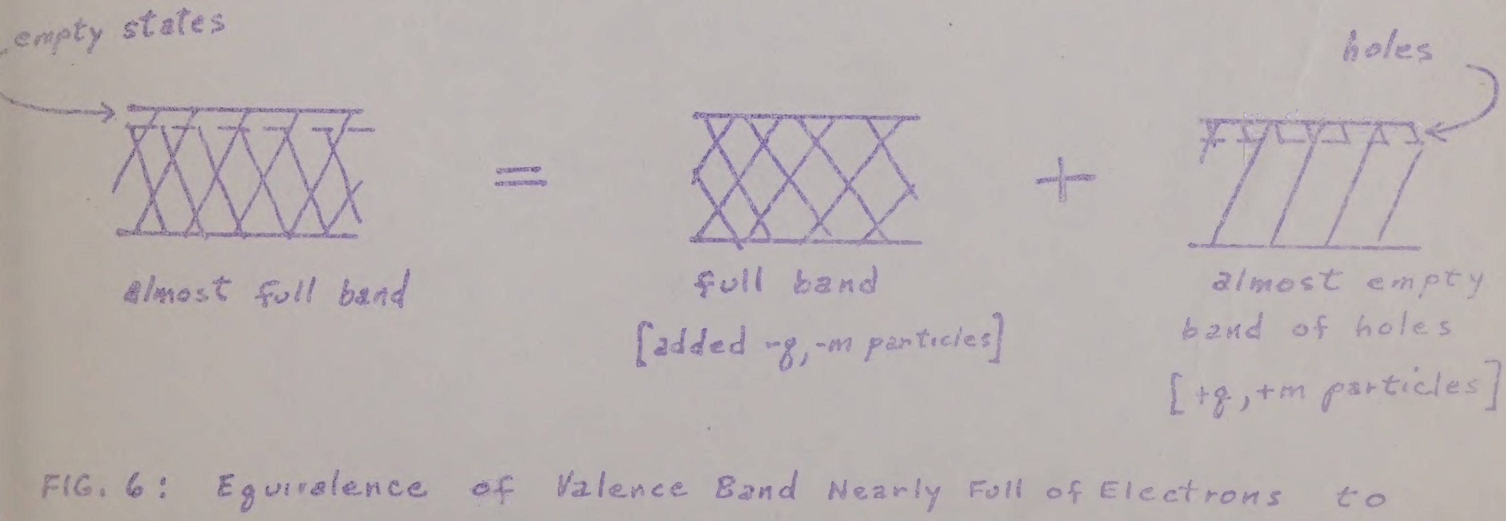


FIG. 6: Equivalence of Valence Band Nearly Full of Electrons to Valence Band Nearly Empty of Holes